Tree-based Bayesian treatment effect analysis

Pedro Henrique Filipini dos Santos

Dissertation presented
to the
Institute of Mathematics and Statistics
of
University of São Paulo
to
obtain the degree
of
Master of Science

Program: Statistics

Advisor: Prof. Hedibert Freitas Lopes, Ph.D.

During the development of this work the author received
financial support from CNPq

São Paulo, February of 2019

Análise de efeitos de tratamento em modelos de árvores Bayesianas

Pedro Henrique Filipini dos Santos

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do Título
de
Mestre em Ciências

Programa: Estatística

Orientador: Prof. Dr. Hedibert Freitas Lopes

Durante o desenvolvimento deste trabalho o autor recebeu auxílo financeiro do CNPq

São Paulo, Fevereiro de 2019

Análise de efeitos de tratamento em modelos de árvores Bayesianas

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 16/05/2019. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Hedibert Freitas Lopes (orientador) - Insper
- Prof. Dr. Rinaldo Artes - Insper
- Prof. Dr. Fabio Gagliardi Cózman - EP-USP

Acknowledgements

To my adviser, Hedibert Freitas Lopes, who believed in my potential when even I had doubted myself, and guided me through moments of adversity.

To my parents, Luiz Carlos dos Santos and Selma Filipini, and my sister, Ana Flávia Filipini dos Santos, for their unconditional support for my studies.

To my godfather and uncle, Rivaldo Almeida de Melo, who is no longer present to see the conclusion of this work, and to my godmother and aunt, Sandra Regina Filipini de Melo, for all the kindness and understanding throughout my life.

To Paula, Karin, Carlos, Erika Monica, Gustavo, Ana Marta, Fernando, Mari-Jô, Luis, Maria Yvette, Carmen, Manuela, Isadora, Aurélio and Lara for the acceptance, trust and support during this difficult phase of my life.

To Rafaela, Renan, Patrícia and Aparecida for supporting me and accepting me in their home as a member of the Coelho family.

To Andreas Anael Pereira Gomes, for the days of late night coding, discussion of models and ideas.

To Lisette Breg, who trusted me and believed in my ability in times of great difficulty.

To my friends of the IME-USP Graduate Program, in particular to the students of room B-154, for all study afternoons, moments of relaxation and emotional support during the last years.

To my friends of USP Athletics and FundUSP, who have always been able to put a smile on my face even in the darkest days.

To friends and colleagues who supported me and kept alive my dream of completing this master's degree.

To the IME-USP teachers, whose teachings I will take for the rest of my life.

To IME-USP staff, who have helped me more times than I can count.

To CNPq, whose financial support made possible the realization of this study.

Agradecimentos

Ao meu orientador, Hedibert Freitas Lopes, que acreditou em meu potencial quando até mesmo eu tinha dúvidas, e me guiou pelos momentos de adversidade.

Aos meus pais, Luiz Carlos dos Santos e Selma Filipini, e minha irmã, Ana Flávia Filipini dos Santos, pelo apoio incondicional aos meus estudos.

Ao meu padrinho e tio, Rivaldo Almeida de Melo, que não está mais presente para ver a conclusão desse trabalho, e à minha madrinha e tia, Sandra Regina Filipini de Melo, por toda a bondade e compreensão ao longo de minha vida.

À Paula, Karin, Carlos, Erika Monica, Gustavo, Ana Marta, Fernando, Mari-Jô, Luis, Maria Yvette, Carmen, Manuela, Isadora, Aurélio e Lara pelo acolhimento, confiança e suporte durante essa fase difícil de minha vida.

À Rafaela, Renan, Patrícia e Aparecida por todo o apoio e por terem me aceitado em seu lar como um membro da família Coelho.

À Andreas Anael Pereira Gomes, pelas diversas madrugadas regadas à programação, discussão de modelos e ideias.

À Lisette Breg, que confiou em mim e acreditou em minha capacidade em momentos de grande dificuldade.

Aos amigos da Pós-Graduação do IME-USP, em especial, aos frequentadores da sala B-154 por todas as tardes de estudo, pelos momentos de descontração e pelo apoio emocional ao longo dos últimos anos.

Aos amigos do Atletismo USP e do FundUSP, que sempre foram capazes de colocar um sorriso em meu rosto até nos dias mais sombrios.

Aos amigos e colegas que me ajudaram a manter vivo o meu sonho de concluir esse mestrado.

Aos professores do IME-USP, cujos ensinamentos eu levarei pelo resto da vida.

Aos funcionários do IME-USP, que me ajudaram mais vezes do que posso contar.

À CNPq, cujo apoio financeiro possibilitou a realização deste estudo.

Abstract

SANTOS, P. H. F. **Tree-Based Bayesian Treatment Effect Analysis** 2019. Dissertation (Master of Science) - Institute of Mathematics and Statistics, São Paulo, 2019.

The inclusion of the propensity score as a covariate in Bayesian regression trees for causal inference can reduce the bias in treatment effect estimations, which occurs due to the regularization-induced confounding phenomenon. This study advocates for the use of the propensity score by evaluating it under a full-Bayesian variable selection setting, and the use of Individual Conditional Expectation Plots as a graphical tool to improve treatment effect analysis. These tools can be used to form groups with different responses to the applied treatment, and to analyze the impact of each variable in the estimated treatment effect.

Keywords: BART; Causality; Propensity Score.

Resumo

SANTOS, P. H. F. **Análise de Efeitos de Tratamentos em Modelos de Árvores Bayesianas** 2019.Dissertação (Mestrado) - Instituto de Matemática e Estatística, São Paulo, 2019.

A inclusão do escore de propensão como uma covariável em modelos de árvores de regressão Bayesianas para inferência causal pode reduzir o viés existente nas estimações de efeitos de tratamento, o qual ocorre devido ao fenômeno de confudimento induzido por regularização. Este estudo defende o uso do escore de propensão por meio de um panorama de seleção de variáveis totalmente Bayesiano, e através do uso de Gráficos de Expectativa Individual Condicional, que se trata de um elemento que pode aprimorar a análise de efeitos de tratamento. Tal ferramental pode ser utilizado como meio de identificar grupos que possuem diferentes respostas ao tratamento aplicado e para analisar o impacto de cada variável no efeito de tratamento estimado.

Palavras-chave: BART; Causalidade; Escore de Propensão.

Contents

Abbreviation

ACF	Autocorrelation Function
BART	Bayesian Additive Regression Trees
BCF	Bayesian Causal Forests
CART	Classification and Regression Trees
CATE	Conditional Average Treatment Effect
DART	Dirichlet Additive Regression Trees
ICE Plots	Individual Conditional Expectation Plots
ITE	Individual Treatment Effect
MCMC	Markov Chain Monte Carlo
OLS	Ordinary Least Squares
PIP	Posterior Inclusion Probability
PDP	Partial Dependence Plot
RIC	Regularization-induced Confounding
RMSE	Root Mean Square Error

List of Symbols

Φ	Standard Normal distribution c.d.f. function
$\|M\|$	The cardinality of set M
$Beta$	Beta Distribution
$\mathcal{D}$	Dirichlet Distribution
$\mathcal{IG}$	Inverse-Gamma Distribution
$\mathcal{N}$	Gaussian Distribution
χ^2	Chi-Square Distribution

List of Figures

List of Tables

Chapter 1

Introduction

1.1 An Approach to Treatment Effect Analysis

When analyzing a dataset the researcher usually wants to find some relationship between some response variable $y = \{y_1, .., y_n\}$ and its covariates $X = \{x_1, ..., x_p\}$, where $x_i = \{x_{i1}, ..., x_{in}\}, i = \{1, ..., p\}$. An usual approach is to fit a linear model that wants to express the relationship between these variables.

Let us consider the following example. Let $i = 1$, $x_{1j} \sim \mathcal{N}(10, 2^2)$, with $j = \{1...n\}$, and $y_j = x_{1j} + \epsilon_j$, where $\epsilon_j \sim \mathcal{N}(0, 1)$. Then, adjusting a linear model to the data gives a line that appears to closely capture the relationship between x_i and y, as seen in Figure 1.1.

However, if the relation between the variables becomes more complex, this kind of technique gives an over-simplistic approach. For example, let us consider that same data from the previous example, but now consider $y_j = \frac{x_{1j} * \sin(x_{1j})}{3} + \epsilon_j$. As seen in Figure 1.2, the data now has different types of relationship as the values of x_{1j} change, and, as expected, the linear model is not able capture the nuances.

Since a more flexible class of models were needed, the so-called nonlinear models began to be more explored by researchers, that ended up developing new techniques to achieve a better understanding or prediction of the response variable y given its covariates X. One of these models is the Bayesian Additive Regression Trees (BART) model (Chipman *et al.*, 2010), which is a "sum-of-trees" model. As seem on Figure 1.3, the BART model easily adapted to the generated data in this example.

Along the many problems that can be approached by nonlinear models, we are going to focus on the problem of estimating Individual Treatment Effects (ITE) and Average Treatment Effects (ATE) in datasets on which individuals are under the influence of an observable binary treatment Z.

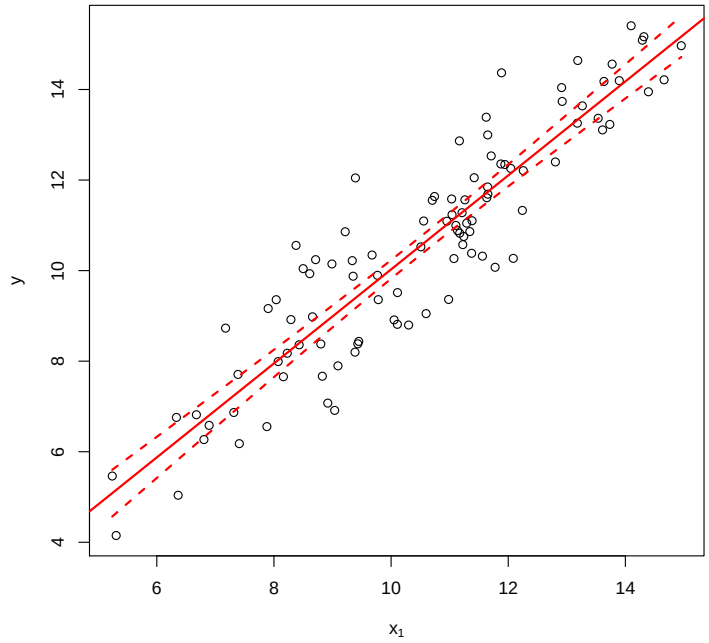

Figure 1.1: *Scatterplot between y and x_1. The solid line is the estimated Ordinary Least Squares linear model, which seems to be a good fit to the data. The dashed lines are the 95% confidence interval estimates.*

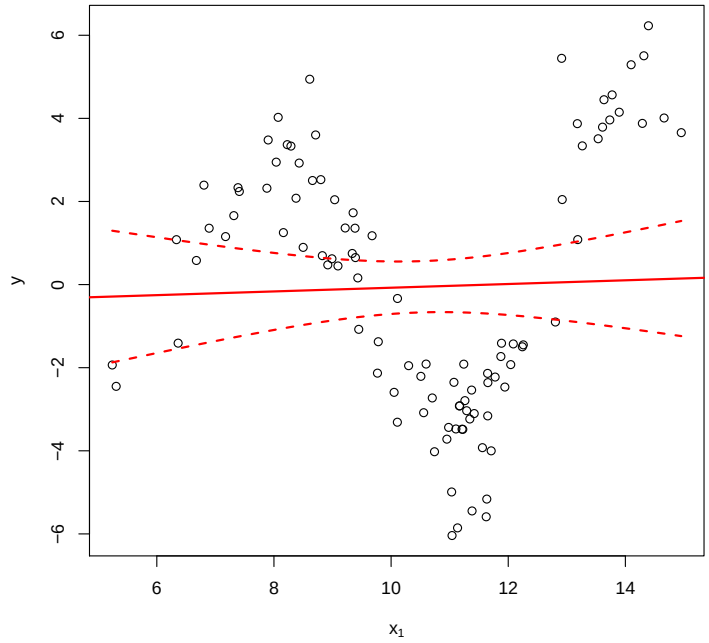

Figure 1.2: *Scatterplot between y and x_1. The solid line is the estimated Ordinary Least Squares linear model, which over-simplifies the relation between the variables. The dashed lines are the 95% confidence interval estimates.*

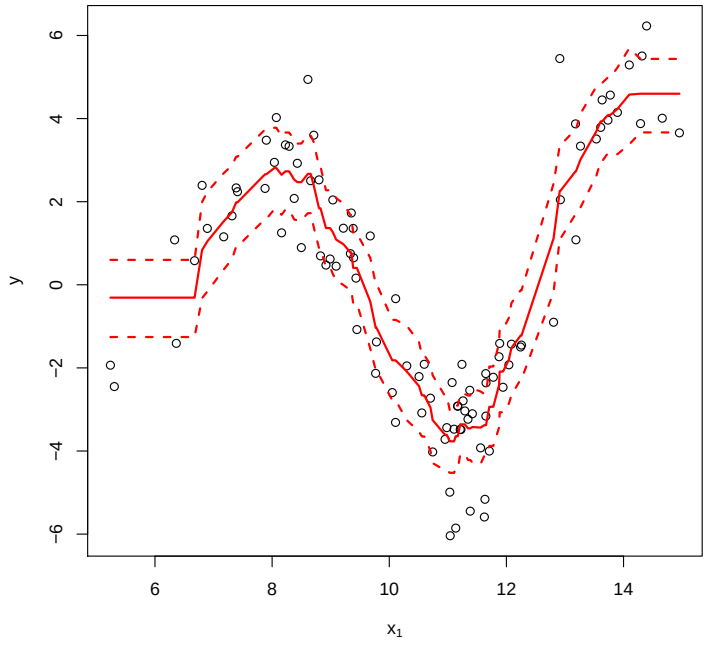

Figure 1.3: *Scatterplot between y and x_1. The solid line is the posterior mean estimates of the BART model (burn-in = 1000; posterior draws = 2000; hypeparameters = default), which seems to be a good fit to the data. The dashed lines are the 95% credible interval estimates.*

This work was motivated by the approaches of Hill (2011) and Hahn *et al.* (2018b). Hill (2011) approach uses the BART model as a tool to estimate Conditional Average Treatment Effects (CATE). Individual Treatment Effects (ITE) estimates are unreliable. Hahn *et al.* (2018b) approach tries to identify treatment effect heterogeneity in a 2-step procedure ("fit-the-fit") via causal BART (BCF) and CART. We introduced a full one-step procedure via causal BART (BCF) with partial dependence plots. In this case, Individual Conditional Expectation (ICE) Plots Goldstein *et al.* (2015) are used to identify treatment effect heterogeneity along with the traditional PDPs Friedman (2001).

More specifically, we use BART model as an alternative to estimate the counterfactual for every single individual in the sample, while including the propensity score, which is the estimated probability of assigning a determined treatment to an individual, as a way to reduce the bias generated by the regularization-induced confounding (RIC) (Hahn *et al.*, 2018a). One of our contributions was the use of Individual Conditional Expectation Plots (ICE Plots) (Goldstein *et al.*, 2015) as a way to perform a sensitivity analysis over the variables and assure that the inclusion of the propensity score reduced the bias in the models. Furthermore, by using Linero (2018) full-Bayesian variable selection, the model was able to identify the correct predictors over the simulations, choosing to include the propensity score between them.

Thus, this work is focused on the methodology that ideally should be followed in order to perform tree-based Bayesian treatment effect analysis on binary treatment data and assure the quality of the results. The codes for the generation of the graphics, simulations and examples used in this dissertation can be accessed on the author‘s GitHub (https://github.com/pedrofilipini/Dissertation).

1.2 Organization of the dissertation

This dissertation is divided in five chapters. Chapter 2 is dedicated to review some of main tree-based models in frequentist and Bayesian literature. Chapter 3 addresses the problem of causality, the main results that are used when dealing with this issue, and some desirable properties when dealing with observational studies. Chapter 4 includes the proposed methodology and tools for analyzing treatment effect when using tree-based Bayesian models, as well as simulations and revisited data analysis in order to corroborate our approach. Finally, Chapter 5 discusses the results, limitations and possible extensions of this dissertation.

Chapter 2

Tree-Based Models

2.1 Introduction

As stated in Hastie *et al.* (2001), tree-based models are based on partitions of the covariate space. Each partition is associated with some kind of function or value, while the rules for the creation of the partitions may vary, as well as the number of trees used in the model.

This class of model has been used as a possible approach to solve regression and classification problems, so this section will be reviewing some of the main models and techniques that were developed in the previous years.

The models that are analyzed in this chapter use binary decision rules to create partitions. That kind of tree model can also be called binary tree.

2.2 Single Tree Models

In order to understand more complex tree-based models it is necessary to know how a single tree model is made. Two models will be reviewed in this section: The Classification and Regression Tree (CART) Model and its Bayesian version, also known as Bayesian CART (B-CART).

2.2.1 The Classification and Regression Tree (CART) Model

The Classification and Regression Tree (CART) model is one of the simplest and intuitive tree-based models. Let y_i be a scalar and represent the response variable for the ith individual, $i = 1, ..., n$, and let $\{\tilde{x}_i = (x_{i1}, ..., x_{ip})\}$ be the vector of p covariates of the ith individual of the sample. The main idea behind this model is that the search for the best point (or split) of the covariate space to partition the data will lead to partitions where the data has reasonable similarity. Each split is given by the rule $\{x_l \leq c\}$ vs $\{x_l > c\}$, $l \in \{1, ..., p\}$, $c \in \mathbb{R}$. An illustration of a binary

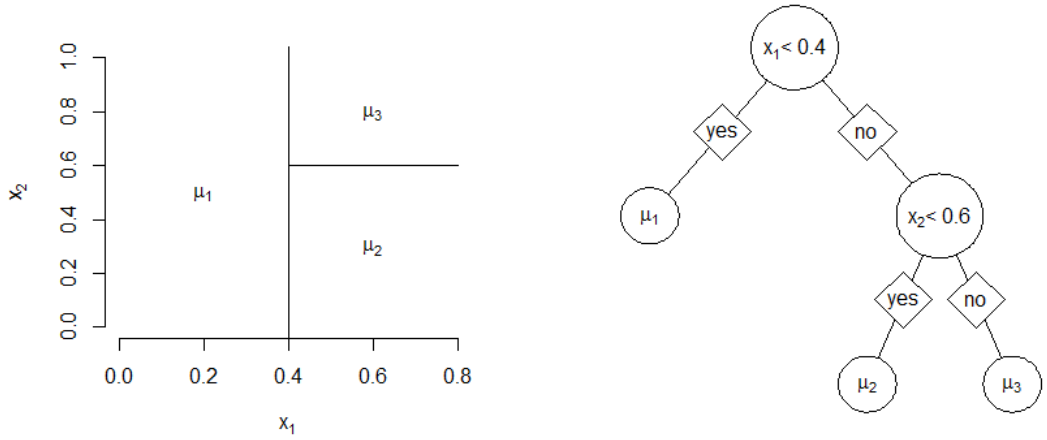

Figure 2.1: *(Left) A partitioned space. (Right) The tree with the corresponding binary splits to the partitioned space. The estimated response for the kth partition is given by* μ_k.

tree with a two splits and three terminal nodes is given in Figure 2.1.

There are many possible ways in which the best split can be found. This section will be based in the algorithms proposed by Breiman *et al.* (1984).

In order to incorporate the split, a tree structure composed of *nodes* will be created. A node represents each split that have been made in the model, while nodes that have no further splits are called *terminal nodes* or *leafs*. Each terminal node should have a value or function associated to it in order to classify the data in that partition. In the CART algorithm the default procedure for regression is to use the mean of the response variable y among the individuals that belong to that specific partition as the response, while for classification, the default procedure is to use the mode of the response variable y among the individuals that belong to that specific partition as the response. Also, since the tree creation relies only on training data, it is easy to use the tree for prediction.

In order to select which variable the split is going to be made, the algorithm performs the steps specified on Algorithm 1. For categorical variables, the process is basically the same, but using the mode of the category as the estimated response of y_i.

One of the attractive resources of the CART model is the possibility of having results that can be considered interpretative. The researcher may want to use the selected splits as a way to draw conclusions about the data that is being studied, but it should be noted that this kind of interpretation can be delusive, since the splits can be generated by spurious relations between the variables, while some important relations may be covered by variables that are predominantly selected as split variable.

Algorithm 1: Algorithm for Splitting a Partition

function Split $(\tilde{y}, \mathbf{x})$;
Input : Training Data with a response vector $\tilde{y}$ and covariate matrix $\mathbf{x} = \{x_1, ..., x_p\}$, where x_i is the vector of the ith covariate with n elements;
Output: Node Splitting Rule for Numeric Variables;
for *(i in* $1 : p$*)* **do**
  Select variable x_i;
  Order the selected variable x_i, where $x_i^{(1)}$ is the minimum value of the sample;
  for *(j in* $1 : (n-1)$*)* **do**
    Set the cutpoint $c_{ij} = \left(x_i^{(j)} + x_i^{(j+1)}\right)/2$, where $x_i^{(j-1)} \leq x_i^{(j)} \leq x_i^{(j+1)}$;
    Set partition P_1 by using the rule $(x_i \leq c_{ij})$ and partition P_2 by using the rule $(x_i > c_{ij})$;
    Calculate $\tilde{Y}_1$ and $\tilde{Y}_2$, where $\tilde{Y}_h$ is the mean of the elements of $\tilde{y} \in P_h$;
    Calculate the MSE for P_1 (MSE_{ij1}) and P_2 (MSE_{ij2});
    Calculate $R_{ij} = (MSE_{ij1} + MSE_{ij2})/2$;
  end
  Calculate $\delta_i = \min\{R_{i1}, ..., R_{i(n-1)}\}$ to select the best split among all possible splits in that variable;
  Let us define $\delta_i = R_{ik_i}$, then c_{ik_i} is the selected split for variable x_i
end
Calculate $Srule = \min\{\delta_1, ..., \delta_p\}$ to select the best variable to split;
Let us define $Srule = \delta_w$, then c_{wk_w} associated with $Srule$ defines the node split rule;
return (Rule: $x_w < c_{wk_w}$)

The CART models suffer several limitations, specially regarding *overfitting*, which is associated with poor data prediction from out-of-sample data which may occur due to the over simplistic design of the model.

Also, since the model need to test every single possible partition in order to take the split, the CART model may be computationally expensive. Furthermore, it is necessary to decide some rule for stopping new splits, as well as rules for pruning large trees.

The split stopping rule is a way to stop the algorithm since the tree may grow until each observation is separated in a different partition, which leads to overfitting. It is usually defined by a minimum number of observations in a terminal node, preventing further splits in that node.

Pruning is an idea that also seeks to avoid overfitting by collapsing some terminal nodes after the tree is created. Usually this allows the tree to increase its prediction performance on out-of-sample data by the cost of losing performance on the training data.

2.2.2 The Bayesian CART

It is also possible to perform a Bayesian approach to fit a tree model such as CART. Two algorithms that were developed in 1998 are often used. The first one was introduced by Chipman *et al.*

(1998), and the second algorithm was introduced by Denison *et al.* (1998). It is important to notice that the two approaches have a considerable number of similarities.

The Chipman *et al.* (1998) approach is the focus of this section. Let y_j be a scalar and represent the response variable for the jth individual, $j = 1, ..., n$, and let $\{\tilde{x_j} = (x_{j1}, ..., x_{jp})\}$ be the vector of p covariates of the jth individual of the sample. Let the tree be formed by a set of binary splits $\{T\}$, and let $\{M = (\mu_1, ..., \mu_b)\}$ define the set of estimated values for the terminal nodes of the tree. Also, each split is of the form $\{x_l < c\}$ vs $\{x_l \geq c\}$, $l \in \{1, ..., p\}$.

Since the model aims to partition the covariate space in relation to the $\{\tilde{y} = (y_1, ..., y_n)\}$, it is assumed that the partitions are independent from each other. Thus, for convenience, it will be assumed that each observation of each partition is independent in relation to the other observations of the same partition. Let us consider that the data from each partition follows a normal distribution. By using these assumptions, it is possible to define the likelihood of $\tilde{y}$ as

$$p\left(\tilde{y} \mid \mathbf{x}, T, M, \sigma\right) = \prod_{i=1}^{b} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_{ij} - \mu_i\right)^2}{2\sigma^2}\right), \tag{2.1}$$

where n_i denotes the number of individuals of the sample that belong to the ith partition, with $i = \{1, 2, ..., b\}$ and $n_1 + n_2 + ... + n_b = n$; y_{ij} denotes the response variable associated with jth individual of the ith partition, $i = \{1, 2, ..., b\}$, $j = \{1, 2, ..., n_i\}$; μ_i denotes the mean parameter associated with the ith partition, $i = \{1, 2, ..., b\}$; and σ^2 denotes the variance parameter of the data.

The variance parameter is considered independent of the other parameters. The priors of the model can be specified as

$$p\left(T, M, \sigma\right) = p\left(M \mid T, \sigma\right) p\left(\sigma \mid T\right) p\left(T\right), \tag{2.2}$$

and

$$p\left(M \mid T, \sigma\right) = \prod_{k=1}^{|M|} p\left(\mu_k \mid T, \sigma\right), \tag{2.3}$$

where $|M|$ is the cardinality of the set M.

The $p\left(\sigma \mid T\right)$ prior is given by

$$\sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu}, \tag{2.4}$$

which is an inverse chi-square distribution. λ and ν are given hyperparameters.

The $p(\mu_k \mid T, \sigma)$ prior is given by

$$\mu_k \mid T, \sigma \sim \mathcal{N}\left(\bar{\mu}, \sigma_\mu^2\right), \tag{2.5}$$

where $\bar{\mu}$ is a given hyperparameter; and $\sigma_\mu^2 = \frac{\sigma^2}{a}$, where a is a given hyperparameter.

The $p(T)$ prior has three parts. The first one is the probability of a split in a determined node, which is given by

$$\mathbb{P}(SPLIT_\gamma) = \eta(1 + d_\gamma)^{-\beta}, \quad \eta \in (0,1), \ \beta \in [0, \infty), \tag{2.6}$$

where η and β are given hyperparameters, and d_γ is the depth of the node γ $\{d \in (0, 1, ...)\}$, where the depth is given by the number of splits between a node and the beginning of the tree.

The second part, which is the probability of selecting a determined variable to perform a split, is given by a discrete Uniform hyperprior. Finally, the third part is given by a discrete Uniform distribution over the possible splits of the variable chosen for the split. These two parts form the rule that is going to be used in each internal node, so they are going to be denoted by

$$\mathbb{P}(RULE_\gamma) = \frac{1}{p_\gamma}\frac{1}{q_{k\gamma}}, \tag{2.7}$$

where p_γ is the total number of available variables to split at node γ; and $q_{k\gamma}$ is the total number of available cutpoints at the node γ for the variable x_k, $\{k = (1, ...p)\}$, which is the selected variable used in the rule.

Now the prior can be defined as

$$p(T) = \prod_{\gamma \in H_{terminal}} (1 - \mathbb{P}(SPLIT_\gamma)) \prod_{\gamma \in H_{internal}} \mathbb{P}(SPLIT_\gamma) \prod_{\gamma \in H_{internal}} \mathbb{P}(RULE_\gamma), \tag{2.8}$$

where $H_{terminal}$ is the set of terminal nodes of the tree T; and $H_{internal}$ is the set of internal nodes of the tree T. Further details over the $p(T)$ prior can be found on Kapelner and Bleich (2016).

Since it may be difficult to deal with varying dimensions over the parameters, Chipman *et al.* (1998) avoided reversible jumps in the covariate space by integrating the set of parameters M out of the likelihood, which is now defined as

$$p\left(\tilde{y} \mid \mathbf{x}, T, \sigma\right) = \int_{M} p\left(\tilde{y} \mid \mathbf{x}, T, \sigma, M\right) p\left(M \mid T, \sigma\right) dM$$

The integral of the likelihood can be simplified to

$$p\left(\tilde{y} \mid \mathbf{x}, T, \sigma\right) = \prod_{i=1}^{b} \left(2\pi\sigma^2\right)^{-\frac{n_i}{2}} \sqrt{\frac{\sigma^2}{n_i\sigma_\mu^2 + \sigma^2}} \exp\left(-\frac{\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2}{2\sigma^2} - \frac{n_i\left(\bar{y}_i - k\right)^2}{2\left(n_i\sigma_\mu^2 + \sigma^2\right)}\right),$$

where $\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$. The simplification is detailed on Appendix A.

To draw the tree samples, Chipman *et al.* (1998) uses a Gibbs Sampler by, first, drawing a tree T with a set of rules by using a Metropolis-Hastings algorithm. Then, the set of parameters M is draw, followed by the draw of the parameter σ.

Chipman *et al.* (1998) introduced four proposals from which the trees can mix by using the Metropolis-Hastings algorithm: GROW (grow a split from a terminal node), PRUNE (collapse a split above two terminal nodes), CHANGE (change a rule from a nonterminal node) and SWAP (swap the rules of a parent and child nonterminal nodes), each one with a fixed probability of being assigned.

As stated before, the draw of M from $p(M \mid T, \tilde{y}, \sigma)$ can be made by a series of independent draws from $p\left(\mu_k \mid T, \tilde{y}, \sigma\right)$, $k = \{(1, ..., |M|)\}$, from the posterior distribution of the parameter

$$\mu_k \mid T, \tilde{y}, \sigma \sim \mathcal{N}\left(\frac{\sigma^2\bar{\mu} + \sigma_\mu^2 n_i \bar{y}_i}{\sigma^2 + \sigma_\mu^2 n_i}, \frac{\sigma^2\sigma_\mu^2}{\sigma^2 + \sigma_\mu^2 n_i}\right).$$

The σ posterior draws from $p(\sigma \mid T, M, \tilde{y}, \mathbf{x})$ are made from an Inverse-Gamma distribution

$$\sigma^2 \mid T, \tilde{y}, M \sim \mathcal{IG}\left(\frac{n+\nu}{2}, \frac{\sum_{i=1}^{b}\sum_{j=1}^{n_i}\left(y_{ij} - \mu_i\right)^2 + \nu\lambda}{2}\right).$$

The intuition behind the B-CART algorithm is that since the Metropolis-Hasting algorithm will be able to identify which of the random splits end up being relevant by using the term $\frac{\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}{2\sigma^2}$ of the likelihood as a way to calculate similarity over the training data. Also, due to the fact that the terminal nodes are only related with the splits by the values sampled from the posterior of σ, the terminal nodes can be interpreted as a separated model for each partition. In general, the Bayesian CART had the issue of its posterior samples being stuck in local modes. This problem was attenuated with the advent of tree ensembles. The model algorithm is given in Algorithm 2.

Algorithm 2: Generating B-CART posterior tree draws

function B-CART $(\tilde{y}, \mathbf{x}, iter)$;

Input : Training Data with a response vector $\tilde{y}$ and covariate matrix $\mathbf{x}$;

Number of MCMC iterations - $iter$;

Output: B-CART posterior tree draws;

Start T as single node trees;

Start M (with $|M| = 1$) filled with zeros;

Start $\sigma = \hat{\sigma}_{OLS}$ as an initial guess;

for *(i in* $1 : iter$*)* **do**

  Select a proposal tree T^* from tree T;

  $T = T^*$ with probability $\alpha(T_j, T^*) = min\left\{\frac{q(T^*,T)}{q(T,T^*)}\frac{p(\tilde{y}|\mathbf{x},T^*,\sigma)p(T^*)}{p(\tilde{y}|\mathbf{x},T_j,\sigma)p(T_j)}, 1\right\}$ or

  $T_j = T$ with probability $1 - \alpha(T_j, T^*)$;

  Draw M from $p(M \mid T, \tilde{y}, \sigma)$;

  Draw σ from $p(\sigma \mid T, M, \tilde{y}, \mathbf{x})$;

  Save the ith posterior draw T, M, σ;

end

2.3 Ensemble of Trees

As stated before, single tree models have some limitations, usually regarding with overfitting or underfitting. Due to the necessity of having more reliable estimations, ensemble of trees were created. Tree ensembles carry some desirable properties that, basically, allow each tree to be a weak learner and capture some specificity of the data. Since the tree ensemble is made in order to assure diversity among the trees, the model end up having a better performance by the cost of losing most of the interpretability of its parameters.

Three kinds of Bayesian tree ensembles will be reviewed. The first is the Bayesian Additive Regression Trees model, which is an extension of the Bayesian CART, the second is the probit-BART, which is a modification of the original BART model to deal with classification problems, and the third is the Bayesian Causal Forests model, which is an BART extension focused on causal inference.

2.3.1 Bayesian Additive Regression Trees (BART) Model

Chipman *et al.* (2010) introduced the BART model, which is a Bayesian nonparametric "sum-

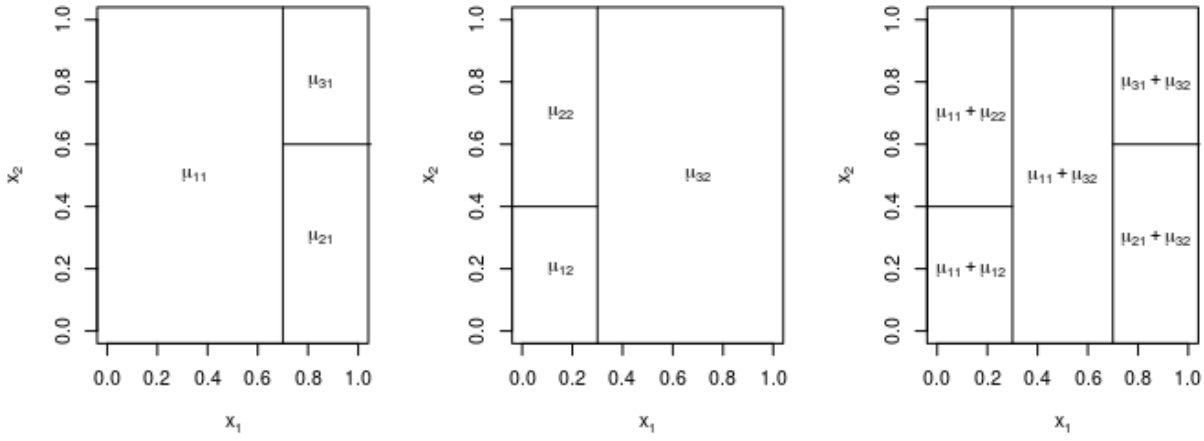

Figure 2.2: *(Left panel) Covariate space partioned by the splits of T_1, with the associate $\mu_{k1}, k = \{1, ..., 3\}$ for every partition. (Middle panel) Covariate space partioned by the splits of T_2, with the associate $\mu_{k2}, k = \{1, ..., 3\}$ for every partition. (Right panel) Covariate space partioned by the sum of T_1 and T_2, with the associate sum of $\mu_{k1} + \mu_{l2}, k = \{1, ..., 3\}, l = \{1, ..., 3\}$ for every partition.*

of-trees" model, where each tree is a weak learner constrained by a regularization prior. Again, the individual tree is formed by a set of binary splits $\{T\}$ from the set of covariates $\{\tilde{x} = (x_1, ..., x_p)\}$, and a set terminal nodes $\{M = (\mu_1, ..., \mu_b)\}$. Each split is of the form $\{x_l < c\}$ vs $\{x_l \geq c\}$ for continuous variables.

The model can be expressed as

$$Y_i = \sum_{j=1}^{m} g\left(\tilde{x}_i; T_j; M_j\right) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right), \tag{2.9}$$

where m is the number of trees on the sum, and $g(\tilde{x}_i; T_j; M_j)$ is the tree function from Chipman *et al.* (1998), which assigns a value $\{\mu_{kj} \in M_j\}$ from the jth tree to $\tilde{x}_i$.

An intuition on how the model works can be achieved by the observation of the behavior of a BART model composed of only two trees. Let us consider that the covariate space is formed of only two variables, $x_1, x_2 \in [0, 1]$, in order for us to identify graphically the splits that were chosen by the model. If we evaluate one single posterior sample of the model, we can consider that the parameters of the model were already chosen, so the set of splits is already defined for the two trees, as well as the sets $M_1 = \{\mu_{11}, \mu_{21}, \mu_{31}\}$ and $M_2 = \{\mu_{12}, \mu_{22}, \mu_{32}\}$. As seen in Figure 2.2, the BART model increases the number of partitions of the covariate space by performing a superposition of splits of each tree, allowing the model to capture more nuances of the relation between the covariates and the response variable, while protecting itself from overfitting by setting a prior that controls the depth of all trees.

The priors of the model can be specified as

$$p\left(\left(T_1, M_1\right), \left(T_2, M_2\right), ..., \left(T_m, M_m\right), \sigma\right) = \left[\prod_{j=1}^{m} p\left(M_j \mid T_j\right) p\left(T_j\right)\right] p\left(\sigma\right), \tag{2.10}$$

and

$$p\left(M_j \mid T_j\right) = \prod_{k=1}^{|M_j|} p\left(\mu_{kj} \mid T_j\right), \tag{2.11}$$

where $|M_j|$ is the cardinality of the set M_j.

The $p\left(\mu_{kj} \mid T_j\right)$ prior works as a regularization prior which constrains each tree to be a "weak learner", which is a model that performs better than chance. For convenience, $\tilde{y}$ is scaled between -0.5 and 0.5, so that the prior given by

$$\mu_{kj} \sim \mathcal{N}\left(0, \sigma_\mu^2\right), \quad \text{with } \sigma_\mu = \frac{0.5}{k\sqrt{m}}, \tag{2.12}$$

helds, for the default setting of $k = 2$, 95% probability that the expected value of the response lies within the interval $(-0.5, 0.5)$. The hyperparameter k can, also, be chosen by cross-validation.

The $p\left(\sigma\right)$ prior is given by

$$\sigma^2 \sim \frac{\nu\lambda}{\chi_\nu^2}, \tag{2.13}$$

which is an inverse chi-square distribution. The hyperparameter λ is given in a way such that $\mathbb{P}\left(\sigma < \hat{\sigma}\right) = q$, where $\hat{\sigma}$ is an initial guess based on the data. Chipman *et al.* (2010) recommends the default setting as $(\nu, q) = (3, 0.90)$.

The $p\left(T_j\right)$ prior is the same of Chipman *et al.* (1998) framework, so the prior for each tree can be defined in the same way as in Equation (2.8).

The posterior of the model can be expressed as

$$p\left(\left(T_1, M_1\right), \left(T_2, M_2\right), ..., \left(T_m, M_m\right), \sigma \mid \tilde{y}, \mathbf{x}\right). \tag{2.14}$$

One way to sample from this posterior distribution is through the Bayesian backfitting algorithm introduced by Hastie and Tibshirani (2000), which, basically, is a Gibbs sampler drawing (T_j, M_j)

with $j \in \{1, 2, ...m\}$, conditionally on $\left(\left(T_{(j)}, M_{(j)}\right), \sigma\right)$, where $\left(T_{(j)}, M_{(j)}\right)$ is the set of $m-1$ trees with its associated terminal node parameters except (T_j, M_j).

To perform a draw from $p\left((T_j, M_j) \mid \left(T_{(j)}, M_{(j)}\right), \sigma\right)$ it is important to notice that $\left(T_{(j)}, M_{(j)}\right)$ have impact on (T_j, M_j) only through

$$\tilde{R}_j \equiv \tilde{y} - \sum_{h \neq j} g\left(\mathbf{x}; T_h; M_h\right), \tag{2.15}$$

in such a way that $p\left((T_j, M_j) \mid \tilde{R}_j, \sigma\right)$ can be sampled using the framework of Chipman *et al.* (1998) for drawing samples of a single tree.

Chipman *et al.* (2010) used the same proposals for the Metropolis-Hastings as Chipman *et al.* (1998): GROW (grow a split from a terminal node), PRUNE (collapse a split above two terminal nodes), CHANGE (change a rule from a nonterminal node) and SWAP (swap the rules of a parent and child nonterminal nodes). Pratola (2016) also presents the ROTATION and PERTURB proposals as an alternative to improve the tree mixing. For further details of proposals GROW, PRUNE and CHANGE, see Kapelner and Bleich (2016). The BART algorithm is given in Algorithm 3.

Following Sparapani *et al.* (2019) instructions, convergence on the BART model can be assessed by the analysis of the trace plots of the σ draws and by the autocorrelation function (ACF), since the draws, ideally, should have no autocorrelation among the σ draws.

Hill (2011) noted that since the BART captures interactions and nonlinearities with ease, handle a large number of covariates that are potential confounders, and have a stable default setup for its priors, so the model is a tool that can be applied to the causal inference setting, specially at the estimation of Conditional Average Treatment Effect in the causal inference setting. Hill (2011) also notes that the model estimates of Individual Treatment Effect have shown great uncertainty.

It must be noted that there is a lack of theoretical proof regarding the efficiency of this class of models, but advances have been made regarding the rate of posterior concentration under some specific parametrization by Ročková and van der Pas (2019) and van der Pas and Ročková (2017).

Algorithm 3: Generating BART posterior tree draws with Bayesian backfitting

function BART $(\tilde{y}, \mathbf{x}, m, iter)$;

Input : Training Data with a response vector $\tilde{y}$ and covariate matrix $\mathbf{x}$;

Number of trees in the ensemble - m;

Number of MCMC iterations - $iter$;

Output: BART posterior tree draws;

Start $T_1, T_2, ..., T_m$ as single node trees;

Start $M_1, M_2, ..., M_m$ (with $|M_j| = 1 \; \forall \; j$) filled with zeros;

Start $\sigma = \hat{\sigma}_{OLS}$ as an initial guess;

for *(i in* $1 : iter$*)* **do**

  for *(j in* $1 : m$*)* **do**

    Calculate $\tilde{R}_j \equiv \tilde{y} - \sum_{h \neq j}^{m} g(\mathbf{x}, T_h, M_h)$;

    Select a proposal tree T^* from tree T_j;

    $T_j = T^*$ with probability $\alpha(T_j, T^*) = min \left\{ \frac{q(T^*, T_j)}{q(T_j, T^*)} \frac{p(\tilde{R}_j | \mathbf{x}, T^*, \sigma) p(T^*)}{p(\tilde{R}_j | \mathbf{x}, T_j, \sigma) p(T_j)}, 1 \right\}$ or

    $T_j = T_j$ with probability $1 - \alpha(T_j, T^*)$;

    Draw M_j from $p(M_j \mid T_j, \tilde{R}_j, \sigma)$;

  end

  Draw σ from $p(\sigma \mid T_1, T_2, ..., T_m, M_1, M_2, ..., M_m, \tilde{y}, \mathbf{x})$;

  Save the ith posterior draw $T_1, T_2, ..., T_m, M_1, M_2, ..., M_m, \sigma$;

end

2.3.2 Probit-BART

An useful extension of the BART model that can be used in classification problems is the probit-BART introduced by Chipman *et al.* (2010). The idea behind the model is straightforward since the model is set as

$$p(\tilde{x}) \equiv \mathbb{P}(Y = 1 \mid \tilde{x}) = \Phi(G(\tilde{x})), \qquad G(\tilde{x}) \equiv \sum_{j=1}^{m} g(\tilde{x}; T_j; M_j), \tag{2.16}$$

where $\Phi(.)$ is the standard normal cumulative distribution function. By using the $\Phi(.)$ function, it is necessary to impose the restriction that $\sigma = 1$.

The probit-BART prior, following the priors (2.10) and (2.11), is given by

$$p\left(\left(T_1, M_1\right), \left(T_2, M_2\right), ..., \left(T_m, M_m\right)\right) = \prod_{j=1}^{m} p\left(M_j \mid T_j\right) p\left(T_j\right), \tag{2.17}$$

and

$$p\left(M_j \mid T_j\right) = \prod_{k=1}^{|M_j|} p\left(\mu_{kj} \mid T_j\right), \tag{2.18}$$

where $|M_j|$ is the cardinality of the set M_j.

The $p\left(T_j\right)$ prior for the jth tree is the same prior used in the B-CART model given in equation (2.8). The $p\left(\mu_{kj} \mid T_j\right)$ prior is defined in a way that wants to ensure that $p\left(\tilde{x}\right)$ remains in regions of high density of the standard normal distribution and that $G\left(\tilde{x}\right)$ shrinks to 0, since $p\left(\tilde{0}\right) = 0.5$. This is done by setting

$$\mu_{kj} \sim \mathcal{N}\left(0, \sigma_\mu^2\right), \quad \text{with } \sigma_\mu = \frac{3}{k\sqrt{m}}, \tag{2.19}$$

where k is a hyperparameter, $k = 2$ being the default. The idea is to concentrate the $G\left(\tilde{x}\right)$ in the interval $(-3, 3)$.

The bayesian backfitting algorithm is also adapted by using Albert and Chib (1993) augmentation idea. Details on the algorithm can be found on Chipman *et al.* (2010). Furthermore, inference can be done exactly the same way as in the BART model.

Since $\sigma = 1$, convergence of the probit-BART model is not as straightforward as in the BART model. Following Sparapani *et al.* (2019) instructions, convergence on the probit-BART model can be made through Geweke convergence diagnostics (Geweke, 1992). In words, the individuals trace plots and ACF functions must be analyzed, as well as the Geweke statistics for each individual, since the distribution of this statistics over the sample must follow a standard normal distribution.

2.3.3 Bayesian Causal Forests (BCF) Model

One possible extension of the BART model focused on the causal inference setting can be achieved by adding linear components to the model (Chipman *et al.*, 2010, p. 295). The Bayesian Causal Forests (BCF) model introduced by Hahn *et al.* (2018b) follows this idea by using a linear combination of two BART models to estimate the value of the response Y_i.

The BCF model is given by

$$Y_i = m\left(\tilde{x}_i, \hat{\pi}\left(\tilde{x}_i\right)\right) + \alpha\left(\tilde{x}_i, \hat{\pi}\left(\tilde{x}_i\right)\right) z_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right), \tag{2.20}$$

where α and m are sum-of-trees functions which are represented as BART models, $\tilde{x}_i$ is vector of p covariates for the ith individual of the sample, and $\hat{\pi}\left(\tilde{x}_i\right)$ is the estimate of the propensity score $\pi\left(\tilde{x}_i\right) = \mathbb{P}\left(Z_i = 1 \mid \tilde{x}_i\right)$ for the ith individual. The propensity score estimate for each individual is introduced in the model as a covariate with the main objective of reducing bias into the estimate of treatment effects. Its role is further explored at Section 4.3.1.

The function m estimates the prognostic effect of each individual, while the function α is used to capture its treatment effect. The BCF has been designed this way because in the original BART setting there was no control on how the model varies in Z, and in this new reparametrization the Z works like an indicator function for the α, enabling the model to aggregate all the covariates interactions regarding treatment effects in the same function.

The functions m and α can have different priors that are adaptable according to its characteristics. Both functions held reasonable results with the default prior settings, but Hahn *et al.* (2018b) have chosen to do two main modifications regarding α. The first modification has been made to support homogeneous treatment effects in the model by two amendments in the T_j prior, more specifically, on Equation (2.6): Considering that the homogeneous treatment effects would be represented by an α function that is a set of trees that are root nodes, the parameter η is adapted to the probability of homogeneous effects $\{\alpha_0 \in (0,1)\}$ by solving $\left\{\alpha_0 = (1-\eta)^{L_\alpha}\right\}$, where L_α is the number of trees in the function α; Setting $\beta = 3$ (instead of $\beta = 2$) to lower the split probability, since the z_i works as an indicator function which can be compared to a first tree split in the variable Z, meaning that all trees in α actually start with depth of 1. The second modification has been made through the implementation of a half-Cauchy hyperprior $\left\{\nu_\alpha \sim \mathcal{C}\left(0, \nu_0\right)_+\right\}$ to the scale parameter of α, where $\alpha\left(\tilde{x}_i\right) \sim \mathcal{N}\left(0, \nu_\alpha\right)$. The default BART uses a constant to its scale parameter, but the change grants a way of avoiding spurious inferences. For further details on the half-Cauchy hyperprior, see Gelman (2006).

Furthermore, the α posterior estimates can be used to analyze the treatment effects at individual level in a way that is possible to identify groups which have positive (or negative) treatment effects within a certain credible interval, allowing a kind of study that was not recommended in Hill (2011) framework due to the lack of estimates robustness.

Chapter 3

Causality

3.1 Introduction

In this study, Rubin (1974) framework will be uses, so, in this case, causality is the study of change in a response variable due to the change in some covariate associated with the response. It is important to note that causality is completely different from correlation, since causality is defined by the order of the events in a sense of cause-and-effect analysis, while correlation simply tries to find some association between these variables.

Due to recent computational advances, some techniques for identifying causal relationships have been developed. The Potential Outcomes framework introduced by Neyman (1923) and Rubin (1974) is the approach that will be used in this study, with some remarks from the the Structural Causal Model (SCM) framework from Pearl (2000).

For this section, the following notation is used: Let us consider a sample of size n. $\tilde{X}_i$ denotes a vector of p covariates $\left\{\tilde{X}_i = (X_{i1}, ..., X_{ip})\right\}$ associated with the ith individual of the sample; Y_i be the scalar response variable associated with the ith individual of the sample; Z_i is the binary treatment variable associated with the ith individual of the sample, where $Z_i \in \{0, 1\}$; the Y_i that has a Z_i realization can be denoted by $Y_i(Z_i)$ and the response variable can be interpreted as $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$.

3.2 Causal Inference

In this work, causality will be studied at the sense of a binary treatment variable. This means that there are two events that will need to be compared, but since only one of these two events can be observed, the other event needs to be estimated. The event that needs to be estimated is known as counterfactual.

The main objective of causal inference is to make conclusions about the changes that occurred in variable of interest and to assess that these changes occurred due to the treatment that is being studied.

3.2.1 Causality on Regression

When dealing with regression, since our variable of interest is numeric, the main objective of causal inference is to estimate the difference between have happened to what would have happened if the other treatment was assigned. This difference will be defined as treatment effect.

However, due to intrinsic individuals particularities, the treatment effect may change. In order to assess that the treatment is effective, it is necessary to account for some level of uncertainty, and to try to separate the population in different groups, in a way that every single individual from a certain group will have similar responses to the assigned treatment. All in all, the main problems are to find ways to define the characteristics of each group and to low the uncertainty related to the estimation of the individual treatment effect.

3.2.2 Observational Studies

Usually there are two kinds of causal inference studies that are conduced. The first kind is the randomized study, which is the gold standard for causal inference. Basically, the randomized study tries to define the groups of control and treatment as homogeneous as possible, in such a way that is possible to make direct comparisons between individuals of those two groups. In this case, it is said that there is sample balance. The main downside of this technique is due to the great cost and difficult to conduct a controlled research. Furthermore, in some areas, like Economics, this kind of research is almost infeasible.

The second kind of study is known as observational study. Basically, the individuals that belong to the treatment and to the control group were not assigned in a homogeneous way, meaning that comparisons between those groups will most likely be biased. In this case, it is said that there is no sample balance. This kind of data is common in most areas where controlled studies are not so easy to conduct.

Usually, when dealing with the causal inference setting, we will be dealing with observational studies, so it is important to use techniques that take the differences between the individuals in each treatment group in account. In randomized studies, the probability of assigning a certain

treatment $\mathbb{P}(Z_i = 1 \mid \tilde{X}_i)$ is known, then

$$Y_i(0), Y_i(1) \perp\!\!\!\perp Z_i \mid \tilde{X}_i, \tag{3.1}$$

holds. However, in observational studies, this is not true and every individual could have a different probability of having the treatment assigned to him. So Rosenbaum and Rubin (1983) define the assumption of *strong ignorability* as a way to allow the analysis of observational studies. This condition require two assumptions:

Assumption 1 - *Unconfoundedness*

$Y(0), Y(1) \perp\!\!\!\perp Z \mid \tilde{X}$

Assumption 2 - *Overlap*

$0 < \mathbb{P}(Z = 1 \mid \tilde{X}) < 1$

The first assumption guarantees that there are no unmeasured confounders in the analysis. As described by Gelman and Hill (2007), *confounders* are covariates that, when not used in the model, produce a misleading estimate of treatment effect, since the effect of the treatment will be confounded with the effect from that specific covariate. One intuition behind this assumption is that Z_i has no impact in the data generation process of $Y_i(Z_i)$, in other words, since the response variable is given by $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ the unconfoundedness assumption is a way to guarantee that both the data generation process of the realization and data generation process of the counterfactual are not affected by the chosen treatment. The second assumption assures that there is a positive probability of assigning each treatment to every individual in the population, always enabling the existence of the counterfactual, thus making it possible to estimate the treatment effect. Also, $\mathbb{P}(Z = 1 \mid \tilde{X})$ is defined as the propensity score.

3.3 The Propensity Score

The propensity score was first introduced as a balancing score, which is an assisting tool to reduce bias in observational studies by balancing the data according to the probability of assigning a treatment to an individual given its vector of covariates. This section is basically a review of the main results from Rosenbaum and Rubin (1983) article.

A balancing score $b(\tilde{x})$ is a function of the covariates $\tilde{x}$ such that

$$\tilde{X} \perp\!\!\!\perp Z \mid b(\tilde{X}), \tag{3.2}$$

which means that, given $b(\tilde{x})$, the distribution of the covariates $\tilde{x}$ is independent of the treatment z, such that, if $z = 1$ or $z = 0$, the distribution of the covariates is the same.

There is a series of theorems and corollaries that guarantee the effectiveness of the methods related with the use of the propensity score. Some of these results are presented here. Further details on these results, as well as most of their proofs, see Rosenbaum and Rubin (1983).

Theorem 1 *Treatment assignment and the observed covariates are conditionally independent given the propensity score, that is*

$$\tilde{X} \perp\!\!\!\perp Z \mid \hat{\pi}(\tilde{X}).$$

Theorem 1 basically states that the propensity score is a balancing score by following the definition of a balancing score.

Theorem 2 *Let $b(\tilde{X})$ be a function of $\tilde{X}$. Then $b(\tilde{X})$ is a balancing score, that is,*

$$\tilde{X} \perp\!\!\!\perp Z \mid b(\tilde{X}),$$

if and only if $b(\tilde{X})$ is finer than $\hat{\pi}(\tilde{X})$ in the sense that $\hat{\pi}(\tilde{X}) = f\left(b(\tilde{X})\right)$ for some function f.

Theorem 2 states that any score that is a function of the propensity score is a balancing score.

Theorem 3 *If treatment assignment is strong ignorable given $\tilde{X}$, then it is strong ignorable given any balancing score $b(\tilde{X})$; that is*

$$Y(0), Y(1) \perp\!\!\!\perp Z \mid \tilde{X},$$

and

$$0 < \mathbb{P}(Z = 1 \mid \tilde{X}) < 1,$$

for all $\tilde{X}$ imply

$$Y(0), Y(1) \perp\!\!\!\perp Z \mid b(\tilde{X}),$$

and

$$0 < \mathbb{P}(Z = 1 \mid b(\tilde{X})) < 1$$

for all $b(\tilde{X})$.

Theorem 3 implies that if *strong ignorability* holds given $\tilde{X}$, then it holds given any balancing score, allowing comparisons in observational studies given a single covariate, which is the balancing score.

Theorem 4 *Suppose treatment assignment is strong ignorable and $b(\tilde{X})$ is a balancing score. Then the expected difference in observed responses to the two treatments at $b(\tilde{X})$ is equal to the average treatment effect at $b(\tilde{X})$, that is,*

$$\mathbb{E}\left(Y\left(1\right) \mid b(\tilde{X}), Z=1\right) - \mathbb{E}\left(Y\left(0\right) \mid b(\tilde{X}), Z=0\right) = \mathbb{E}\left(Y\left(1\right) - Y\left(0\right) \mid b(\tilde{X})\right).$$

Theorem 4 is the main result from Rosenbaum and Rubin (1983), since it states that, under *strong ignorability*, individuals with the same value of $b(\tilde{X})$, but not necessarily the same values of $\tilde{X}$, and with different values of Z, can be treated as the counterfactuals of each other, allowing the estimation of the average treatment effect. In words, under *strong ignorability*, the propensity score do a kind of partitioning of the individuals of the sample in such a way that individuals from the same partition can be compared in order to obtain an estimation for their treatment effect.

By using these theorems, Rosenbaum and Rubin (1983) also introduced matched sampling, subclassification, and covariance adjustment techniques by using the propensity score. The use of these techniques is corroborated by three corollaries.

Corollary 1 (Pair Matching on Balancing Scores) *Suppose treatment assignment is strong ignorable. Further suppose that a value of a balancing score $b(\tilde{X})$ is randomly sampled from the population of units, and then one treated, $Z = 1$, unit and one control, $Z = 0$, unit are sampled with this value of $b(\tilde{X})$. Then the expected difference in response to the two treatments for the units in the matched pair equals the average treatment effect at $b(\tilde{X})$. Moreover, the mean of matched pair differences obtained by this two-step sampling process is unbiased for the average treatment effect.*

Corollary 2 (Subclassification on Balancing Scores) *Suppose treatment assignment is strong ignorable. Suppose further that a group of units is sampled using $b(\tilde{X})$ such that:*

1. *$b(\tilde{X})$ is constant for all units in the group, and*
2. *at least one unit in the group received each treatment.*

Then, for these units, the expected difference in treatment means equal the average treatment effect at that value of $b(\tilde{X})$. Moreover, the weighted average of such differences, that is, the directly

adjusted difference, is unbiased for the treatment effect, when the weights equal to the fraction of the population at $b(\tilde{X})$.

Corollary 3 (Covariance Adjustment on Balancing Scores) *Suppose treatment assignment is strong ignorable, so that in particular,* $\mathbb{E}\left(Y(z) \mid b(\tilde{X}), Z = z\right) = \mathbb{E}\left(Y(z) \mid b(\tilde{X})\right)$ *for balancing score* $b(\tilde{X})$. *Further suppose that the conditional expectation of* $Y(z)$ *given* $b(\tilde{X})$ *is linear:*

$$\mathbb{E}\left(Y(z) \mid b(\tilde{X}), Z = z\right) = \alpha_z + \beta_z b(\tilde{X}), \qquad (z = 0, 1).$$

Then the estimator

$$(\hat{\alpha}_1 - \hat{\alpha}_0) + \left(\hat{\beta}_1 - \hat{\beta}_0\right) b(\tilde{X}),$$

is conditionally unbiased given $b(\tilde{X}_i)$, $(i = 1, ..., n)$, *for the treatment effect at* $b(\tilde{X})$, *namely* $\mathbb{E}\left(Y(1) - Y(0) \mid b(\tilde{X})\right)$, *if* $\hat{\alpha}_z$ *and* $\hat{\beta}_z$ *are conditionally unbiased estimators of* α_z *and* β_z, *such as least squares estimators. Moreover,*

$$(\hat{\alpha}_1 - \hat{\alpha}_0) + \left(\hat{\beta}_1 - \hat{\beta}_0\right) \bar{b},$$

where $\bar{b} = \frac{\sum_{i=1}^{n} b(\tilde{X}_i)}{n}$, *is unbiased for the average treatment effect if the units in the study are a simple random sample of the population.*

These techniques are widely used for treatment effect estimation, but since this section is focused on the properties of the propensity score, they are not detailed in this study. For more details on matched sampling, subclassification, and covariance adjustment, see Rosenbaum and Rubin (1983).

It is important to notice that these results hold for the true values of the balancing scores, including the propensity score, which cannot be observed. Since the propensity score needs to be estimated, it is essential to assure that strong ignorability holds.

Theorem 5 *Suppose* $0 < \hat{\pi}(\tilde{X}) < 1$. *Then*

$$prop\left(Z = 0, \tilde{X} = \tilde{a} \mid \hat{\pi}(\tilde{X}) = \hat{\pi}(\tilde{a})\right) = prop\left(Z = 0 \mid \hat{\pi}(\tilde{X}) = \hat{\pi}(\tilde{X})\right) prop\left(\tilde{X} = \tilde{a} \mid \hat{\pi}(\tilde{Z}) = \hat{\pi}(\tilde{a})\right),$$

where $prop\left(Z = 0, \tilde{X} = \tilde{a} \mid \hat{\pi}(\tilde{X}) = \hat{\pi}(\tilde{a})\right)$ *is the proportion of the* n *units with* $Z = 0$ *and* $\tilde{X} = \tilde{a}$ *among all units with* $\hat{\pi}(\tilde{X}) = \hat{\pi}(\tilde{a})$; $prop\left(Z = 0 \mid \hat{\pi}(\tilde{X}) = \hat{\pi}(\tilde{X})\right)$ *is the proportion of the* n *units with* $Z = 0$ *among all units with* $\hat{\pi}(\tilde{X}) = \hat{\pi}(\tilde{X})$; *and* $prop\left(\tilde{X} = \tilde{a} \mid \hat{\pi}(\tilde{Z}) = \hat{\pi}(\tilde{a})\right)$ *is the proportion*

of the n units with $\tilde{X} = \tilde{a}$ *among all units with* $\hat{\pi}(\tilde{Z}) = \hat{\pi}(\tilde{a})$.

Corollary 4 *Suppose the n units (considering a sample of size n) are a random sample from an infinite population, and suppose* $\tilde{X}$ *takes on only fintely many values in the population and at each such value* $0 < \pi(\tilde{X}) < 1$. *Then with probability 1 as* $n \to \infty$, *subclassification on* $\hat{\pi}(\tilde{X})$ *produces sample balance, that is, Theorem 5 holds.*

Theorem 5 and Corollary 4 state that sample estimates of the propensity score can be used to induce sample balance in the sample.

3.4 A Remark about the Assumption of Strong Ignorability

Even though the propensity score has been widely used in many fields as a way to estimate treatment effects in observational studies, Pearl (2009) discuss about the controversy of using or not the propensity score methods in order to estimate the treatment effect. The main reason for this controversy is that practitioners may end up assuming that *strong ignorability* holds, or even simply as many covariates as possible (Pearl, 2009, p. 350).

It is important to remember that currently it is not possible to assess when *strong ignorability* holds or not, and since by the time that this study was conducted there were no results known by the authors that could actually replace the assumption of *strong ignorability*, the results presented are only guaranteed to hold in simulations since the researcher know the data generation process and can ensure *strong ignorability*.

Furthermore, Pearl (2009) argues that the propensity score can be seen by the SCM framework and that conditioning on many covariates might even increase the bias of the treatment effect estimation. This problem is caused due to the effect of unmeasured *confounders*.

In words, if *strong ignorability* does not hold, the results estimated by these methods might be deceiving, specially due to unmeasured *confounders*. Further development in this area is needed, maybe with the use of graphs and techniques introduced by the SCM framework.

Chapter 4

Treatment Effect Analysis

4.1 Introduction

The advent of Bayesian computation on the previous decades allowed the creation of models with a high degree of complexity, while Monte Carlo Markov Chains (MCMC) algorithms allow the estimation of models that were previously considered infeasible. One of these models is the Bayesian Additive Regression Trees (BART) (Chipman *et al.*, 2010).

Hill (2011) applied the BART models to the causal inference setting, more specifically on the estimations of binary treatment effects for observational studies, and the results were promising. These models can be affected by the regularization-induced confounding (Hahn *et al.*, 2018a), which states that, in the presence of confounding, the regularization of these models may lead to biased estimations of the treatment effects. Hahn *et al.* (2018b) argue that, under strong ignorability, this problem can be eased through the use of the propensity score (Rosenbaum and Rubin, 1983) among the covariates of the model.

This section has two main contributions. The first is the application of a sensitivity analysis in the causal inference setting through the use of Individual Conditional Expectation (ICE) Plots (Goldstein *et al.*, 2015). The second is to corroborate the inclusion of the propensity score on Bayesian regression tree models (Hahn *et al.*, 2018b) by using simulations and the full-Bayesian variable selection proposed by Linero (2018).

Section 4.2 introduces notation and a revision on tree-based Bayesian regression trees for causal inference. Section 4.3 specify how the propensity score and the Individual Conditional Expectation (ICE) Plots can be used to properly perform treatment effect analysis on Bayesian regression trees. Section 4.5 have simulations on which the techniques advocated by this study are used. Finally in Section 4.6 a real data analysis from Hahn *et al.* (2018b) is revisited using the framework presented in this Section.

4.2 Treatment Effect Estimation with Bayesian Regression Trees Models

Capital roman letters are used to denote random variables, while realizations are denoted in lower case. Vectors are denoted by a tilde on the top of the variable, and matrices are denoted by bold variables. Let Y denote a scalar response, Z denote a binary treatment effect and $\tilde{X}$ denote a vector of p covariates $\left\{\tilde{X} = (X_1, ..., X_p)\right\}$. In such way that the triplet $\left(Y_i, Z_i, \tilde{X}_i\right)$ denotes the observation of the ith individual of a sample size n.

Following the notation of Imbens (2004), the Y_i that has a Z_i realization can be denoted by $Y_i(Z_i)$ and the response variable can be interpreted as $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$, where $Z_i \in \{0, 1\}$. It is important to notice that only $Y_i(0)$ or $Y_i(1)$ can be observed, while the unobserved unit is called counterfactual. For this framework *strong ignorability* (Rosenbaum and Rubin, 1983) is assumed to hold, which means that Assumption 1 and Assumption 2 are assumed to hold.

Using the framework adopted by Hahn *et al.* (2018b) for expressing treatment effects, it follows that the Individual Treatment Effect (ITE) can be represented as

$$\alpha\left(\tilde{x}_i\right) = \mathrm{E}\left(Y_i \mid \tilde{x}_i, Z_i = 1\right) - \mathrm{E}\left(Y_i \mid \tilde{x}_i, Z_i = 0\right), \tag{4.1}$$

where $\alpha\left(\tilde{x}_i\right)$ is the ITE for the ith individual of the sample.

As in Hill (2011), the Conditional Average Treatment Effect (CATE) is defined by

$$\frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{E}\left(Y_i(1) \mid \tilde{x}_i\right) - \mathrm{E}\left(Y_i(0) \mid \tilde{x}_i\right)\right], \tag{4.2}$$

which is equivalent to the average of the individual treatment effects of the sample.

To assess the simulations in this chapter two measures are used. The first is the CATE RMSE defined as

$$\sqrt{\frac{1}{L}\sum_{l=1}^{L}(\hat{\gamma}^{(l)} - CATE)^2}, \tag{4.3}$$

where $\hat{\gamma}^{(l)}$ is the CATE estimated in the lth posterior draw, $l = (1, ..., L)$.

The second is the ITE RMSE, which is defined as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\alpha}(\tilde{x}_i) - \left[\mathrm{E}\left(Y_i(1) - Y_i(0) \mid \tilde{x}_i\right)\right]\right)^2}, \tag{4.4}$$

where $\hat{\alpha}(\tilde{x}_i)$ is the ITE estimated through the posterior mean. Basically, it is an RMSE since the true treatment effect for each individual is known when dealing with simulations.

4.3 Treatment Effect Analysis

The BART model presented in Section 2.3.1 can be used to estimate the ITE by

$$\hat{\alpha}\left(\tilde{x}_i\right) = \hat{f}\left(\tilde{x}_i, 1\right) - \hat{f}\left(\tilde{x}_i, 0\right), \tag{4.5}$$

where $\hat{f}\left(\tilde{x}_i, z_i\right)$ consist of the posterior mean from the prediction of the estimated model.

In the case of the BCF model from Section 2.3.3, as the α function is estimated separately from the prognostic effect, the BART prior of α already gives the posterior draws from the estimated model as an output. It should be noted that since both models outputs are given by posterior draws, it is straightforward to construct credible intervals for the estimated treatment effects with the use of quantiles from these draws.

The following subsections introduce some methods and tools to assist the analysis of treatment effects in these models.

4.3.1 The RIC and the Role of the Propensity Score

The term "regularization-induced confounding" (RIC) was introduced by Hahn *et al.* (2018a) into the setting of linear models with homogeneous treatment effects and further expanded by Hahn *et al.* (2018b) to the BART models, which despite having a good predictive performance, have shown biased treatment effect estimation and lack of robustness at the individual level estimates when applied to the causal settings, as noted by Hill (2011). Hahn *et al.* (2018b) tries to avoid the RIC phenomenon by including the estimate of the propensity score as a covariate in the BART and the BCF models.

As shown by Hahn *et al.* (2018b), the inclusion of the propensity score allows the tree-based models to adapt more easily in cases where the data exhibits complex confounding. Since the propensity score naturally simplifies the number of required splits in a context of parsimonious trees, it allows the model to focus on other interactions between the variables, reducing bias and improving the predictions of the model.

Under the sparsity setting it is possible to assess the use of the propensity score in the model by the variable selection framework introduced by Linero (2018) in its Dirichlet Additive Regres-

sion Trees (DART) models, which assigns a sparsity-inducing Dirichlet hyperprior (instead of an Uniform hyperprior) on the probability of choosing a variable to split on, which is given by

$$(s_1, ...s_P) \sim \mathcal{D}\left(\frac{\theta}{P}, \frac{\theta}{P}, ..., \frac{\theta}{P}\right),$$

where P is the number of covariates in the model and θ is given by the prior

$$\frac{\theta}{\theta + \rho} \sim Beta\,(a, b)\,,$$

with $a = 0.5$, $b = 1$ and $\rho = P$ in this study.

Linero (2018) suggests two approaches to perform full-Bayesian variable selection: The first one is the analysis of the posterior draws from the Dirichlet hyperprior and the second one is to use the method of variable selection proposed by Barbieri and Berger (2004) and calculate the Posterior Inclusion Probability (PIP) for each variable.

The former approach is straightforward, since the posterior draws allow the construction of credible intervals for the probability of choosing any variable. The latter approach can be performed by simple verifying, for each iteration of the MCMC, if the variable x_l, $l \in \{1, ..., p\}$, is used in at least one splitting rule of the tree ensemble. If it is used, the value 1 is assigned, and, if not, 0 is assigned. The PIP_l is the mean of these indicator functions for variable x_l over all iterations of the MCMC, and the variable x_l will be selected if $PIP_l > 0.5$.

4.3.2 A Visualization Tool: ICE Plots

Friedman (2001) introduced the Partial Dependence Plot (PDP) in order to allow the visualization of the impact that a variable have in the response, performing a kind of sensitivity analysis.

The partial dependence function for the ith observation is defined as

$$\hat{f}\,(x_{Si}) = \frac{1}{n}\sum_{j=1}^{n} \hat{f}\,(x_{Si}, \tilde{x}_{Cj})\,, \tag{4.6}$$

where n is the sample size, $\hat{f}(.)$ is the prediction function from the model, $S \subset \{1, ..., p\}$ is the index of one of variables of interest, C is a subset such that $C \cup S = \{1, ..., p\}$ and $C \cap S = \emptyset$, x_{Si} is a scalar from the variable of interest from the ith observation of the training data, and $\tilde{x}_{Cj}$ is the vector of covariates from the subset C of the observation j from the training data. The curve is made from the partial dependence functions of all observations from $\tilde{x}_S$. In general, $S \subset \{1, ..., p\}$

with $|S| \geq 1$, but there are no means of plotting the PDP for $|S| > 2$.

Chipman *et al.* (2010) suggest the use of PDP to analyze the marginal effect of the variables in relation to the response. It must be noted that the implementation of this technique in BART models actually give draws from the posterior distribution, thus, it is easy to acquire the credible intervals for the PDP curve.

This tools have already been applied to BART models in the treatment effect setting by Green and Kern (2012), where the authors create curves regarding each variable in relation to the CATE, which takes the place of the response variable. This is possible due to a slight modification in the algorithm by using

$$\hat{f}_{CATE}\left(x_{Si}\right) = \frac{1}{n}\sum_{j=1}^{n}\left[\hat{f}\left(x_{Si}, \tilde{x}_{Cj}, z_i = 1\right) - \hat{f}\left(x_{Si}, \tilde{x}_{Cj}, z_i = 0\right)\right]. \tag{4.7}$$

But since this tool only evaluates the mean of the response, it cannot be used into the analysis of individual treatment effects.

Goldstein *et al.* (2015) introduced the ICE Plots by noting that since the PDP curve could conceal heterogeneous effects in the response variable, it was necessary a tool for analyzing the marginal effect at each individual. By using the same notation of Equation (4.6) the ICE function for the ith individual is defined by

$$\hat{f}\left(x_{Sij}\right) = \hat{f}\left(x_{Si}, \tilde{x}_{Cj}\right), \tag{4.8}$$

where the ICE curve for the jth individual is formed by calculating the ICE function for every i of the sample. This way, there will be n curves in the plot, and the PDP curve can be obtained by averaging the ICE curves.

The main reason for the development of such a tool is due to the difficulty to interpret parameters in machine learning methods, more specifically, the models known as "black box" models, such as Support Vector Machines, Random Forests, Neural Networks, etc.

This analysis can be very useful in the context of treatment effects, since in the setting of heterogeneous treatment effects it is possible to look for indications of groups that have different reactions to the applied treatment. Like in Equation (4.7), the ICE curves can be adapted to the ITE setting by estimating

$$\hat{f}_{ITE}\left(x_{Sij}\right) = \left[\hat{f}\left(x_{Si}, \tilde{x}_{Cj}, z_i = 1\right) - \hat{f}\left(x_{Si}, \tilde{x}_{Cj}, z_i = 0\right)\right]. \tag{4.9}$$

instead of using the formula described in Equation (4.8).

The ICE Plots have two interesting features regarding the response of each individual. The first is the centered ICE (c-ICE) plot which removes any level difference between the individuals, thus allowing a simple graphical of heterogeneity among the individuals. The c-ICE function for the ith individual can be defined as

$$\hat{f}_{cent}\left(x_{Sij}\right) = \hat{f}\left(x_{Sij}\right) - \mathbf{1}\hat{f}\left(x^*, x_{Ci}\right). \tag{4.10}$$

where x^* is a fixed location in the range of x_S.

The second feature is the derivative ICE (d-ICE) plot, which presents an estimation of the partial derivative of the ICE curve in relation to x_S. According to Goldstein *et al.* (2015), the d-ICE plot intuition is that the observation of different values for the d-ICE function at the same point suggest the existence of interactions between the variable that is being analyzed and the remaining covariates of the model. This graphic generated by the *ICEbox* package also plots an estimation of the standard deviation among the points, which is an indicator of heterogeneity among the individuals. Further details on the d-ICE can be found on Goldstein *et al.* (2015).

As a way to illustrate the interpretation of these plots, the toy example from Goldstein *et al.* (2015) is replicated. Consider the following data generation process:

$$Y_i = 0.2x_{i1} - 5x_{i2} + 10x_{i2}\mathbf{1}\left(x_{i3} \geq 0\right) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1), \tag{4.11}$$

$$X_{i1}, X_{i2}, X_{i3} \sim \mathcal{U}\left(-1, 1\right). \tag{4.12}$$

where $i = 1, ..., 100$. Depending on the value of X_3, the coefficient from X2 can be 5 or -5. This relation can be seen of Figure 4.1.

A BART model (burn-in = 2000; posterior size = 1000; thinning = 100) with default priors was used as an estimator of the ICE function. The model seems to have achieved convergence since the trace plot for σ apparently traverse the sample space adequately (Figure 4.2) and the ACF function indicate that there is low autocorrelation (Figure 4.3).

The ICE Plots were generated by using the *ICEbox* package. For variable x_1, the ICE Plot on Figure 4.4 seems to show almost have almost no impact in the model. Figure 4.7 seems to suggest that x_1 have a slight positive impact in general, but it is inconclusive. The d-ICE plot (Figure 4.10) do not seem to show any interactions between x_1 and other covariates. It was expected this sort of behavior of x_1 on the ICE Plots, since the impact of the variable might be covered by the error

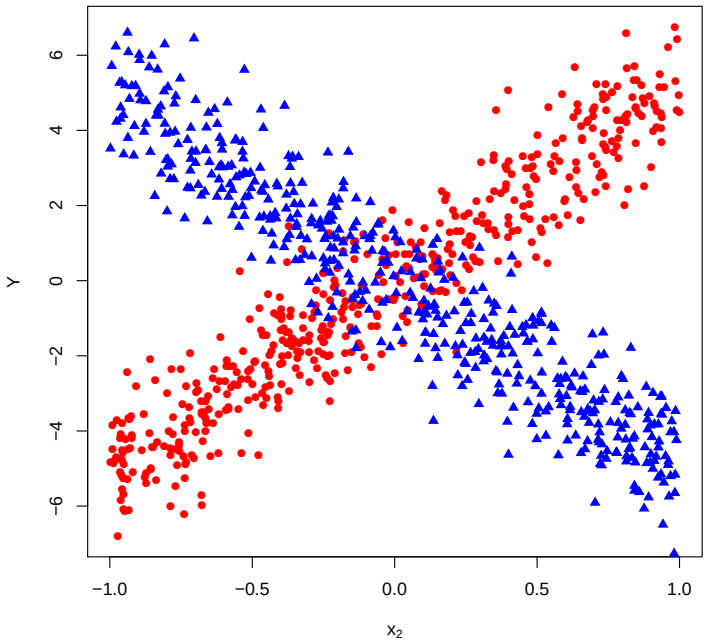

Figure 4.1: *ICE Plot Example - x_2 and y. The blue triangles represent observations where $x_3 < 0$, while the red circles represent individuals where $x_3 \geq 0$.*

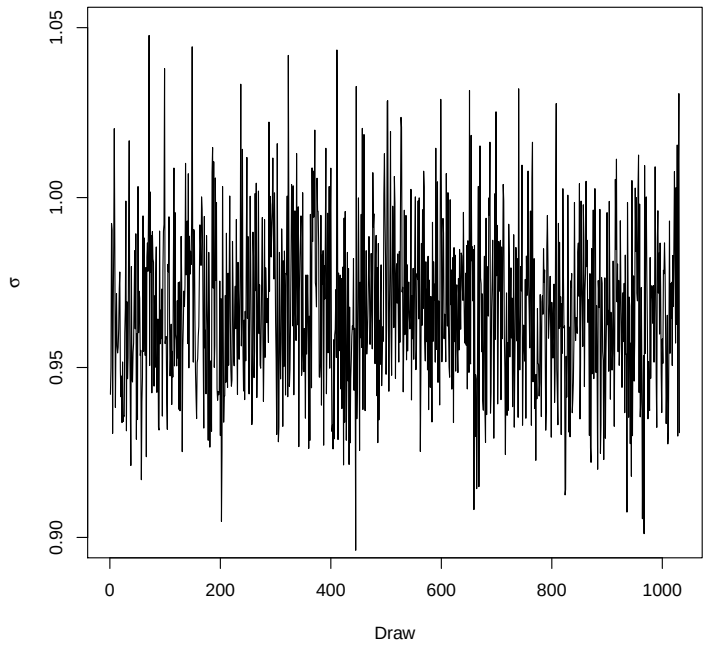

Figure 4.2: *BART model (ICE Plot Example) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

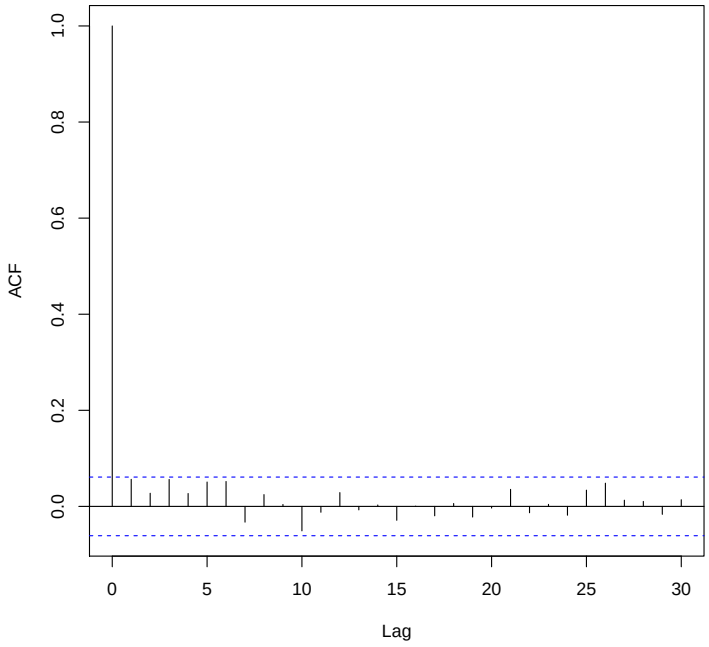

Figure 4.3: *BART model (ICE Plot Example) - ACF function for the σ draws. Apparently there is low autocorrelation among the draws.*

term added to the model.

The ICE plot on Figure 4.5 shows that the PDP curve may be misleading since it would give the researcher an indicative that there is no relation among the response and X_2. The c-ICE plot on Figure 4.8 illustrates the heterogeneity among the data, showing that some individuals have a negative effect as X_2 increases, while others have a positive effect. Finally, the d-ICE plot suggests that since the variance is high among the whole sample, there might be an interaction among the covariates of the model. For x_2 the ICE plots apparently managed to capture the relations of the data generation process.

The ICE Plot related to x_3 (Figure 4.6) is not so clear, but looking at the c-ICE Plot (Figure 4.9) it is possible to notice that when x_3 is bigger than 0, then an effect that might be between -10 and 10 is added. Figure 4.12 indicates that might be an interaction between x_2 and some other covariate when x_3 is near 0. For x_3 the ICE plots apparently managed to capture the relations of the data generation process.

4.4 Assessing the Use of the Propensity Score - A Toy Example

This toy example serves as one more way to advocate over the use of the propensity score as a covariate in tree-based Bayesian models for causal inference. This example is based on Hahn *et al.*

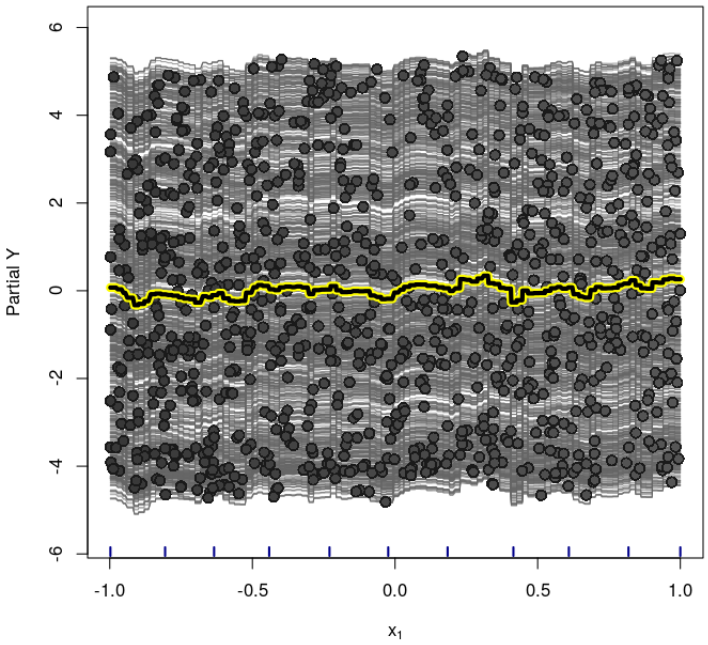

Figure 4.4: *ICE Plot Example - ICE Plot for variable* x_1.

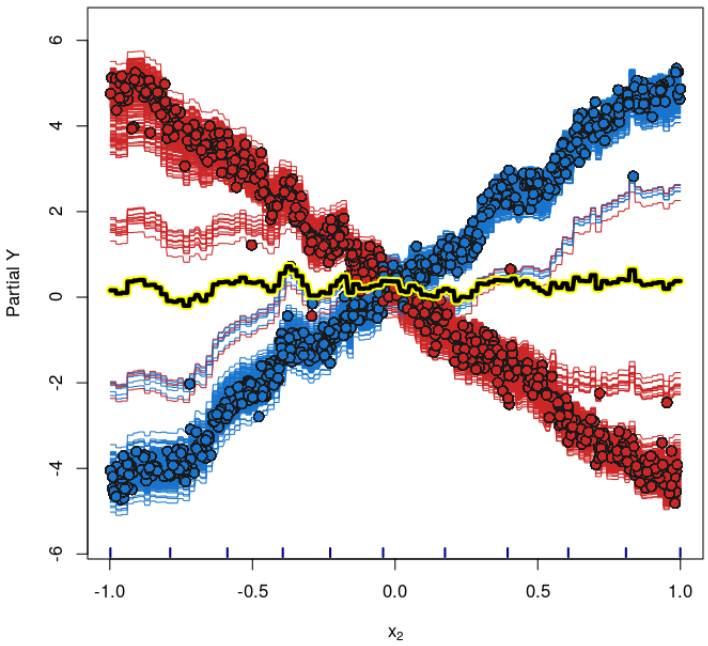

Figure 4.5: *ICE Plot Example - ICE Plot for variable* x_2. *For the observations in blue* $x_3 < 0$, *while for the observations in red* $x_3 \geq 0$.

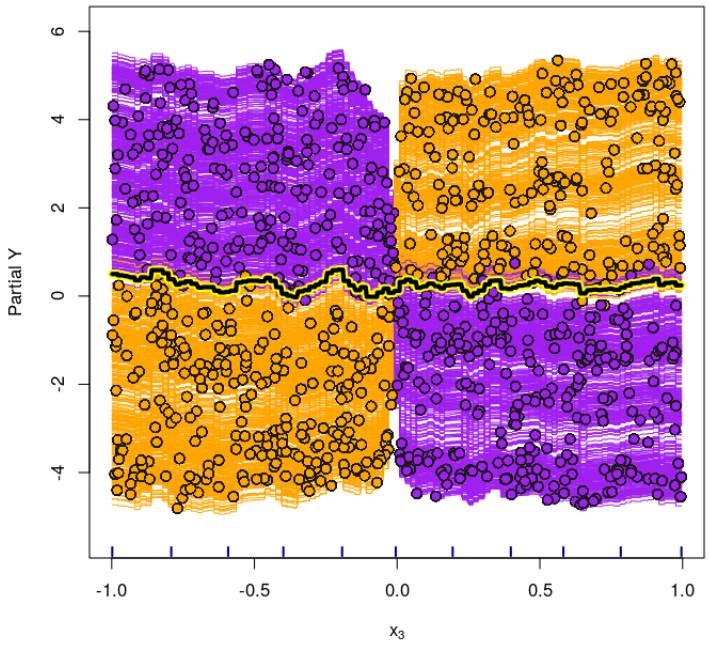

Figure 4.6: *ICE Plot Example - ICE Plot for variable x_3. For the observations in purple $x_2 < 0$, while for the observations in orange $x_2 \geq 0$.*

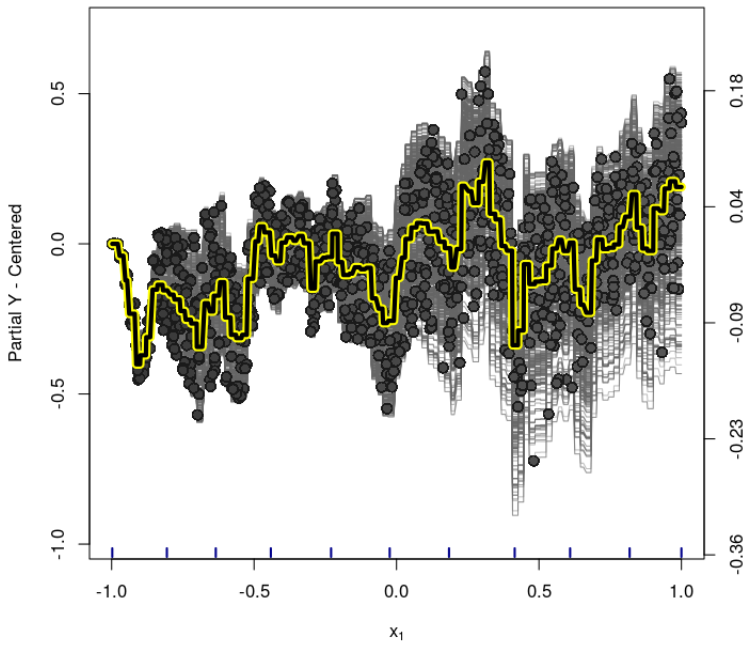

Figure 4.7: *ICE Plot Example - Centered ICE Plot for variable x_1.*

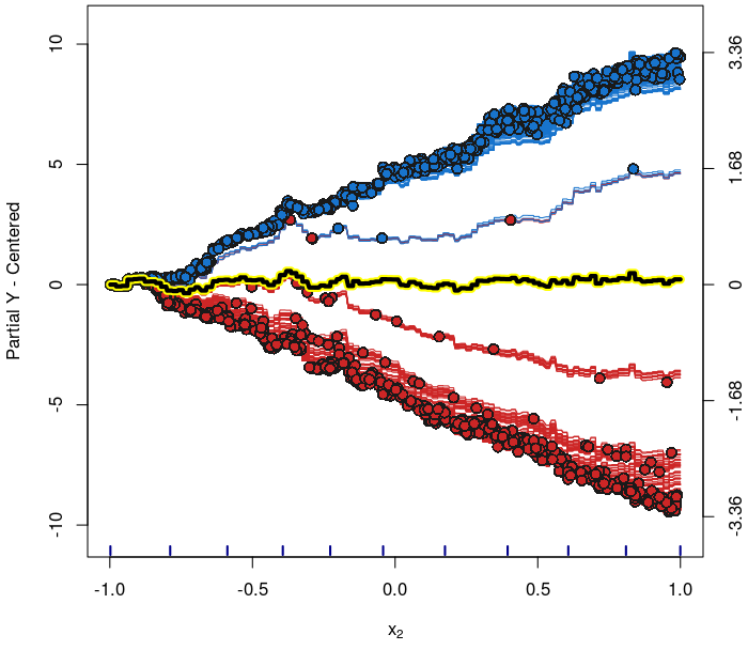

Figure 4.8: *ICE Plot Example - Centered ICE Plot for variable x_2. For the observations in blue $x_3 < 0$, while for the observations in red $x_3 \geq 0$.*

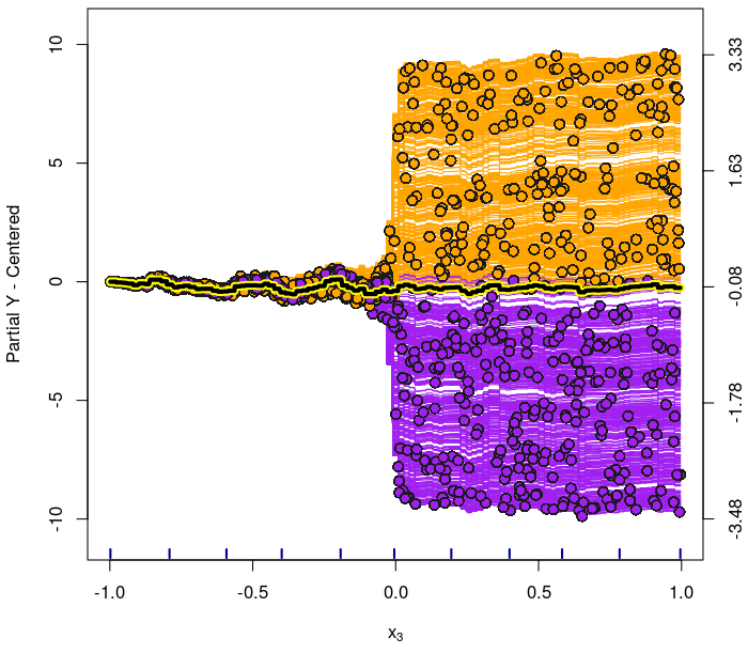

Figure 4.9: *ICE Plot Example - Centered ICE Plot for variable x_3. For the observations in purple $x_2 < 0$, while for the observations in orange $x_2 \geq 0$.*

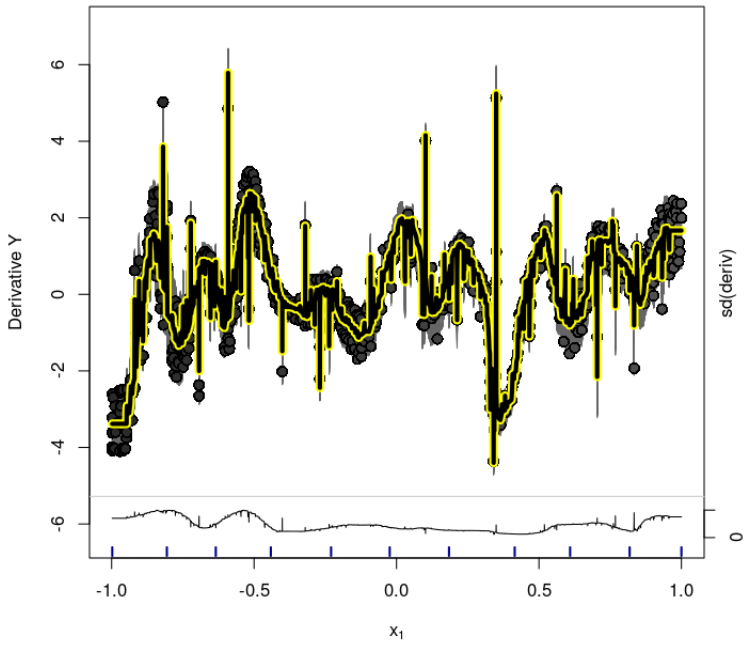

Figure 4.10: *ICE Plot Example - Derivative ICE Plot for variable* x_1.

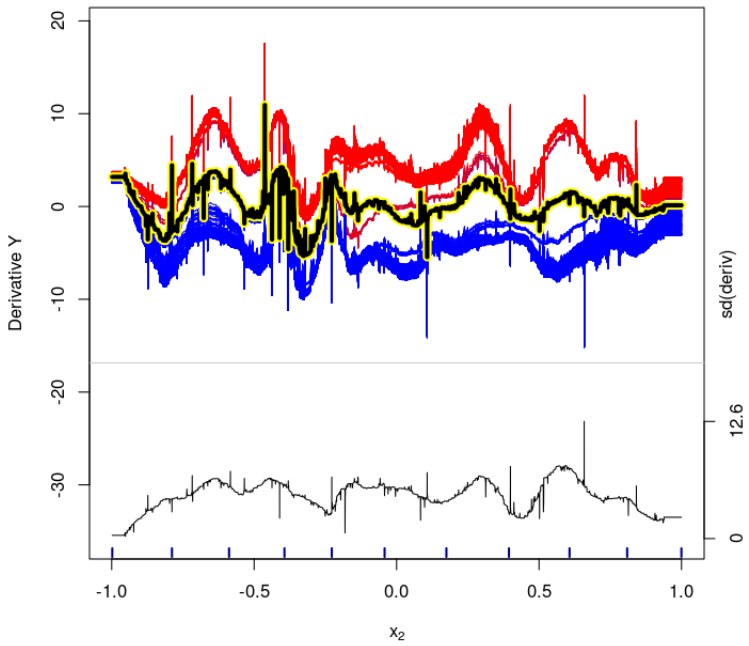

Figure 4.11: *ICE Plot Example - Derivative ICE Plot for variable* x_2. *For the observations in blue* $x_3 < 0$, *while for the observations in red* $x_3 \geq 0$.

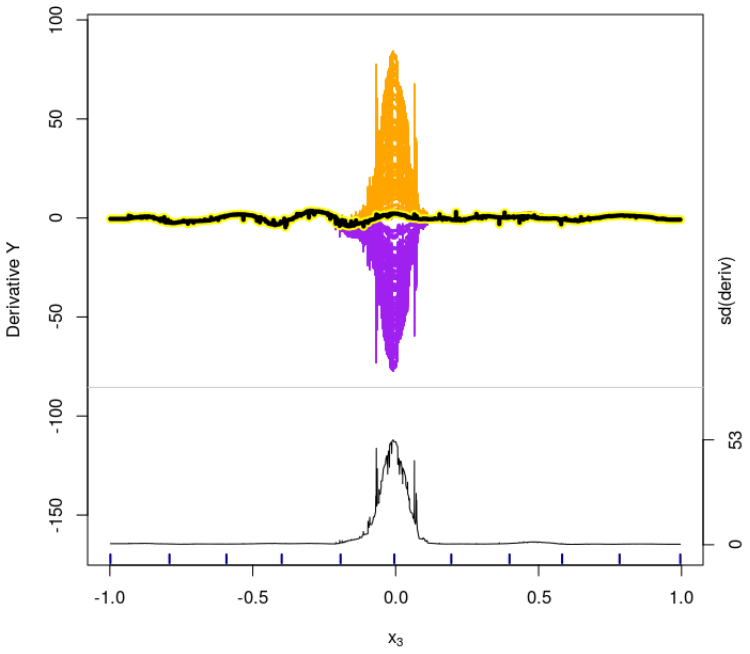

Figure 4.12: *ICE Plot Example - Derivative ICE Plot for variable x_3. For the observations in purple $x_2 < 0$, while for the observations in orange $x_2 \geq 0$.*

(2018b) toy example. Let us consider $n = 200$, where the index i represents the ith individual of the sample, $i = (1, ..., 200)$. The data is generated as

$$X_{i1}, X_{i2}, X_{i3} \sim \mathcal{N}(0, 1).$$

$$\mu_i = \mathbf{1}(x_{i1} < x_{i2}) - \mathbf{1}(x_{i1} \geq x_{i2}),$$

$$P(Z_i = 1 \mid x_{i1}, x_{i2}, x_{i3}) = \Phi(\mu_i),$$

$$\alpha_i = 0.5 * \mathbf{1}(x_{i3} > -3/4) + 0.25 * \mathbf{1}(x_{i3} > 0) + 0.25 * \mathbf{1}(x_{i3} > 3/4).$$

$$Y_i = \mu_i + Z_i \alpha_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, 0.5^2).$$

where $\Phi(.)$ is the standard normal cumulative distribution function.

The idea behind this example is that the rule for μ_i is difficult to be estimated by a tree model, since it represents a diagonal cut in the covariate space betweeen x_1 and x_2. However, since the propensity score is different for every value of μ_i, if the tree chooses to do a split by using the propensity score as a covariate, it would be the same to perform a diagonal cut in the covariate space between x_1 and x_2. The variable x_3 defines the intensity of the treatment effect for each individual. The true $CATE$ is defined as the mean of the true treatment effect of the individuals

of the sample.

The following models were analyzed:

- Vanilla: Y_i estimated by a BART model using $\tilde{x}_i$ as covariates;
- Oracle (Oracle-BCF): Y_i estimated by a BART (BCF) model using $\tilde{x}_i$ and $\pi(\tilde{x}_i)$, the true value of the propensity score, as covariates;
- PS-BART (PS-BCF): Y_i estimated by a BART (BCF) model using $\tilde{x}_i$ and $\hat{\pi}(\tilde{x}_i)$, estimated by the posterior mean of the probit-BART, as covariates;
- GLM-BART (GLM-BCF): Y_i estimated by a BART (BCF) model using $\tilde{x}_i$ and $\hat{\pi}(\tilde{x}_i)$, estimated by GLM (which is considered as a naive approach to estimate the propensity score), as covariates;
- Rand-BART (Rand-BCF): Y_i estimated by a BART (BCF) model using $\tilde{x}_i$ and $\hat{\pi}(\tilde{x}_i)$, given by a random Uniform distribution, as covariates.

The default prior settings were used for every model. The burn-in is set as 5000, the posterior size as 500 and the thinning used for this example is 150.

As expected, the models that used an estimate of the propensity score as a covariate seemed to have exhibited a reduction on the CATE bias. This is shown by this posterior CATE estimates boxplots for each model on Figure 4.13. It is important to notice that the Oracle and the Oracle-BCF models are infeasible since on real data analysis it is not possible to use the true propensity score as a covariate.

The trace plots of the σ draws and the autocorrelation function (ACF) graphics for these models are supplied in Appendix B.1 to support convergence. Apparently, the trace plots from all models showed almost no autocorrelation and seem to traverse the sample space adequately. These are suggestions that the models have achieved convergence.

The convergence of the probit-BART estimates is analyzed following Sparapani *et al.* (2019) instructions. The ACF plot in Figure 4.14 shows that the individuals have almost no autocorrelation for the first 16 lags. The trace plot in Figure 4.15 seem to indicates that the response of the individuals traverse the sample space adequately. Furthemore, 96.5% of the Geweke statistics calculated for the whole sample are within the 95% range in Figure 4.16, which suggest that convergence apparently have been achieved.

The ICE Plots for this example are supplied on Appendix B.2. As expected, the Oracle models show almost no bias on the average and individual treatment effect estimates due to the inclusion

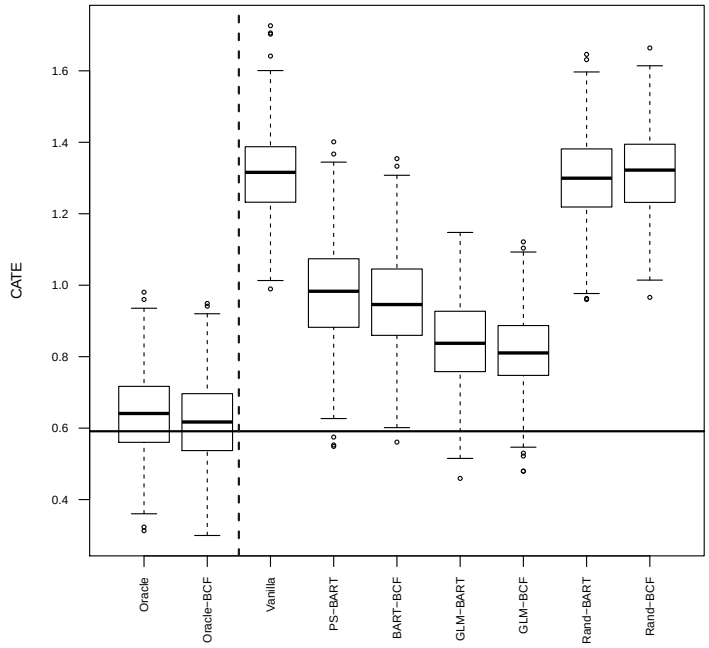

Figure 4.13: *Toy Example - Boxplots of CATE estimates over 500 posterior draws for every model. The models to the left of the dashed vertical line use the true propensity score as a covariate, thus those models are infeasible. The horizontal solid line is the true CATE for the sample.*

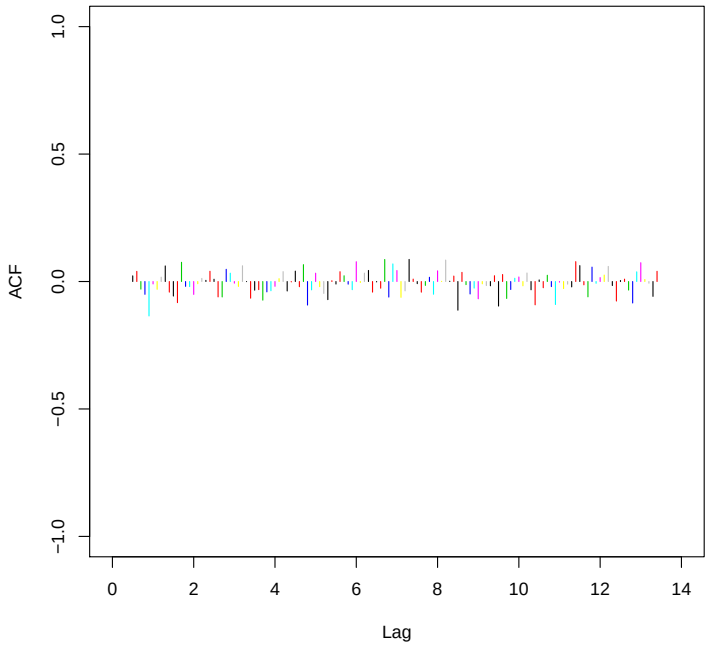

Figure 4.14: *Probit-BART (Toy Example) - ACF functions of 10 sampled individuals. Apparently, the autocorrelation among the 500 posterior draws is low in the sample.*

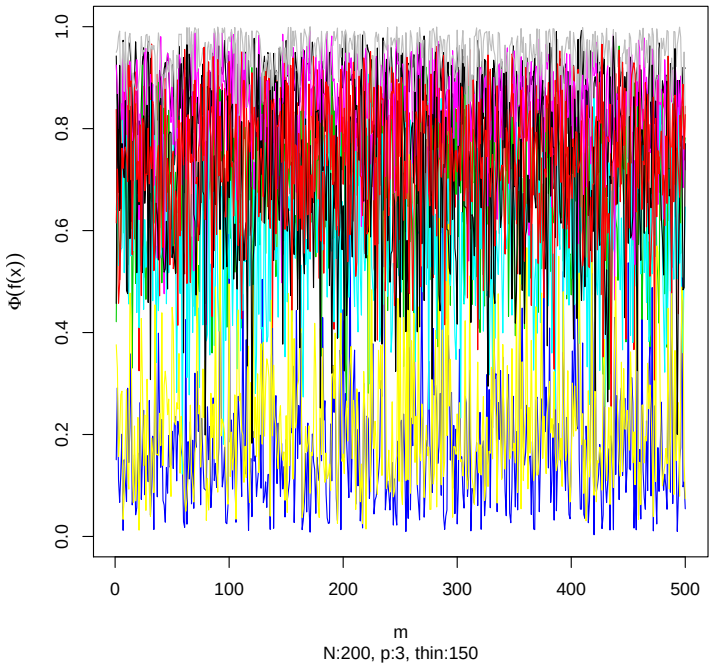

Figure 4.15: *Probit-BART (Toy Example) - Trace plots of 10 sampled individuals. Apparently, the traces traverse the sample space adequately among the 500 posterior draws.*

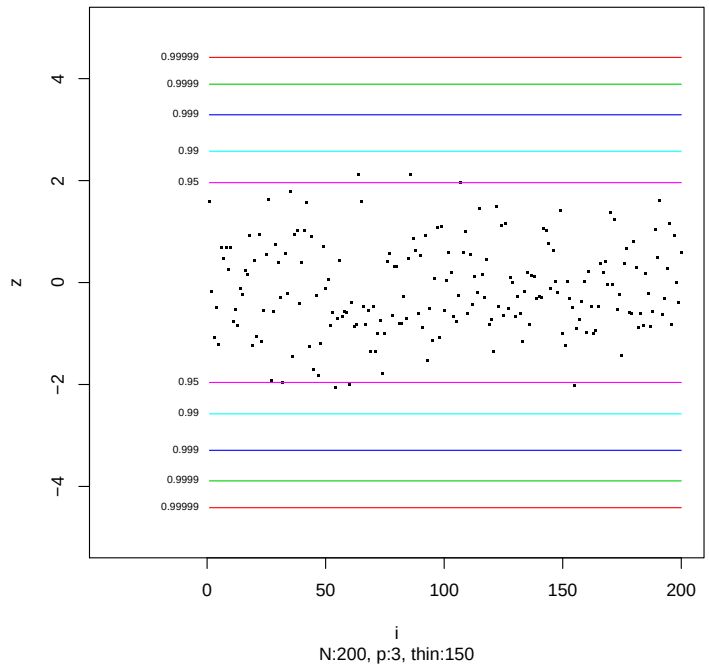

Figure 4.16: *Probit-BART (Toy Example) - Geweke statistics for each individual of the sample. Apparently convergence has been attained since 96.5% of the sample observations are within the 95% range.*

Model	Mean	Standard Deviation	Maximum	Minimum
Oracle	138.324	28.826	390.748	90.794
Vanilla	171.714	45.484	488.433	104.233

Table 4.1: *Toy example - Mean, Standard Deviation, Maximum and Minimum of the measure $\kappa(\epsilon)$ for 500 posterior draws in each BART Model.*

of the propensity score. The PS-BART and the GLM-BART models have shown a reduction on bias since for all covariates the true CATE mostly remained inside the 95% credible interval for the PDP. The PS-BCF and the GLM-BCF also have shown bias reduction over the CATE, but the uncertainty related to the propensity score estimates lead to wider credible intervals for the PDP. The Rand models, as expected, performed poorly due to the inclusion of an irrelevant covariate and the lack of a propensity score estimate. Since there was no interaction among the covariates in the data generation process, the c-ICE Plots and the d-ICE Plots individual curves should be in the same place as the PDP curve, which usually happens in this example.

Ročková and van der Pas (2019) introduced the $\kappa(\epsilon)$ as a measure that would be relevant to compare tree-based models, including BART models. Basically a matrix $A(\epsilon)$ filled with 0s and 1s that maps the whole partition space of the model is created. The measure $\kappa(\epsilon)$ is given by $\lambda_{max}(\epsilon)/\lambda_{min}(\epsilon)$ where $\lambda_{max}(\epsilon)$ ($\lambda_{min}(\epsilon)$) is the maximum (minimum) singular value of $A(\epsilon)$. In words, the $\kappa(\epsilon)$ allows a tree-based model to be compared in relation to the redundancy of its partitions since the existence of redundant trees in the ensemble will make $\kappa(\epsilon)$ to be large.

In this example the $\kappa(\epsilon)$ was calculated for the Vanilla and Oracle models for every posterior draw as seem on Figure 4.17. As expected, the $\kappa(\epsilon)$ for the Oracle model had the minimum, maximum, median, mean and standard deviation measures lower than the Vanilla model, as seem on Table 4.1, which is a clear indication that the Oracle model held partitions less redundant among its trees than the Vanilla model.

Since the BCF models uses a set of trees dedicated to the treatment effect estimation, these sets of trees were analyzed by using the $\kappa(\epsilon)$. The results are on Figure 4.18 and Table 4.2. Basically, the Oracle BCF model presented the lowest mean among the analyzed models, but a standard deviation slightest higher than the PS-BCF model. All the models seemed to have similar performances, which is an indicative that the BCF models have trees that, in general, are not redundant.

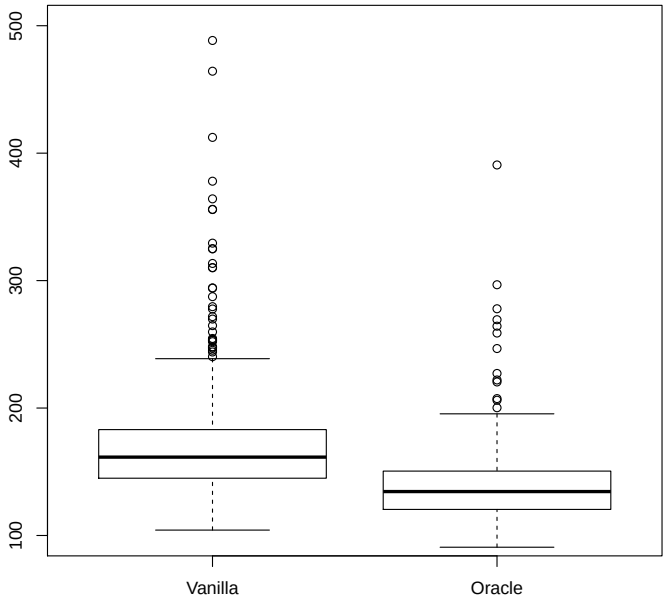

Figure 4.17: *Toy example - Boxplots of $\kappa(\epsilon)$ for 500 posterior draws in each BART Model. The Oracle Model uses the true propensity score as a covariate, thus it is infeasible.*

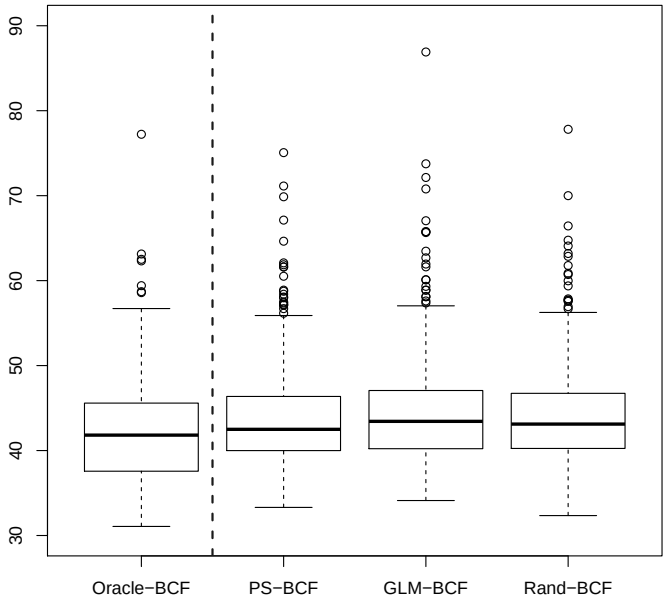

Figure 4.18: *Toy example - Boxplots of $\kappa(\epsilon)$ for 500 posterior draws of α in each BCF Model. The Oracle-BCF Model, which is left to the dotted vertical line, uses the true propensity score as a covariate, thus it is infeasible.*

Model	Mean	Standard Deviation	Maximum	Minimum
Oracle-BCF	42.175	6.145	77.227	31.068
PS-BCF	43.814	5.933	75.066	33.307
GLM-BCF	44.422	6.301	86.911	34.117
Rand-BCF	44.055	5.626	77.813	32.343

Table 4.2: *Toy example - Mean, Standard Deviation, Maximum and Minimum of the measure* $\kappa(\epsilon)$ *for 500 posterior draws of* α *in each BCF Model.*

4.5 Simulations

In order to assess the tools that have been proposed in Section 4.3, some simulations were performed. In Section 4.5.1 the simulations were made in order to corroborate the use of the propensity score as a covariate in BART models by using the ICE Plots to evaluate the ITE of the sample. In Section 4.5.2 the simulations were performed in the sparsity setting, to evaluate how often the propensity score is used by the model in relation to the other variables. Only the BART models were studied at Section 4.5.2, but these methods can be extended to the BCF models as well. Furthermore, in Section 4.5.3 the treatment effect estimates of the models used in the simulations were evaluated.

Zigler and Dominici (2014) pointed that since the propensity score carries uncertainty about its estimations, so it is natural to advocate the use of a Bayesian framework in it. Following Hahn *et al.* (2018b), the probit-BART (Chipman *et al.*, 2010) posterior mean is used as the propensity score estimate for each individual in the simulations. As a comparative, the frequentist approach of estimating the propensity score by Generalized Linear Models with *logit* link was used.

All calculations were performed in R version 3.4.4 (R Core Team, 2017) by using the *packages* BART version 2.2 (McCulloch *et al.*, 2019), bcf version 1.2.1 (Hahn *et al.*, 2017) (this version was slightly modified to allow prediction) and ICEbox version 1.1.2 (Goldstein *et al.*, 2015). As in Hill (2011), the priors, hyperparameters, and hyperpriors were held under default setting in this study, but cross-validation can be performed in order to improve results.

4.5.1 Simulation Based on Real Data

In this scenario, the following models were analyzed:

- Vanilla: Y_i estimated by a BART model using $\tilde{x}_i$ as covariates;
- Oracle (Oracle-BCF): Y_i estimated by a BART (BCF) model using $\tilde{x}_i$ and $\pi(\tilde{x}_i)$, the true value of the propensity score, as covariates;

- PS-BART (PS-BCF): Y_i estimated by a BART (BCF) model using $\tilde{x}_i$ and $\hat{\pi}(\tilde{x}_i)$, estimated by the posterior mean of the probit-BART, as covariates;
- GLM-BART (GLM-BCF): Y_i estimated by a BART (BCF) model using $\tilde{x}_i$ and $\hat{\pi}(\tilde{x}_i)$, estimated by GLM (which is considered as a naive approach to estimate the propensity score), as covariates;
- Rand-BART (Rand-BCF): Y_i estimated by a BART (BCF) model using $\tilde{x}_i$ and $\hat{\pi}(\tilde{x}_i)$, given by a random Uniform distribution, as covariates.

Those simulations were replicated 1000 times in order to assess the results. In each model the first 50000 draws from the MCMC were treated as burn in, while the posterior draws had size 1000. The thinning used for the probit-BART was 500, while the other models had a thinning of 100.

Infant Health and Development Program (IHDP) Dataset

This simulation is based on Hill (2011), and uses the Infant Health and Development Program (IHDP) dataset ($n = 985$). In order to simplify the simulation, only the the following covariates are used: birth weight (x_1); head circunference (x_2); weeks born preterm (x_3); birth order (x_4); neonatal health index (x_5) and age of the mother (x_6). The response surface was generated by

$$Y_i = \beta_1 x_{i1} + ... + \beta_6 x_{i6} + \mu_i + Z_i \alpha_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, 0.5^2).$$

The predictors were standardized for data generation, and the β_i's were sampled from (0, 1, 2, 3, 4) with probabilities (0.05, 0.1, 0.15, 0.2, 0.5).

The true propensity score and the real treatment effects were generated based on Hahn *et al.* (2018b) example as it follows,

$$\mu_i = \mathbf{1}(x_{i1} < x_{i2}) - \mathbf{1}(x_{i1} \geq x_{i2}),$$

$$P(Z_i = 1 \mid x_{i1}, ..., x_{i6}) = \Phi(\mu_i),$$

$$\alpha_i = 0.5 * \mathbf{1}(x_{i3} > -3/4) + 0.25 * \mathbf{1}(x_{i3} > 0) + 0.25 * \mathbf{1}(x_{i3} > 3/4).$$

The Figure 4.19 is composed by the boxplots of posterior CATE for each model for the one replication of the simulation, but it is important to point out that similar results were found for the other replications. The horizontal line is the real CATE for this specific iteration.

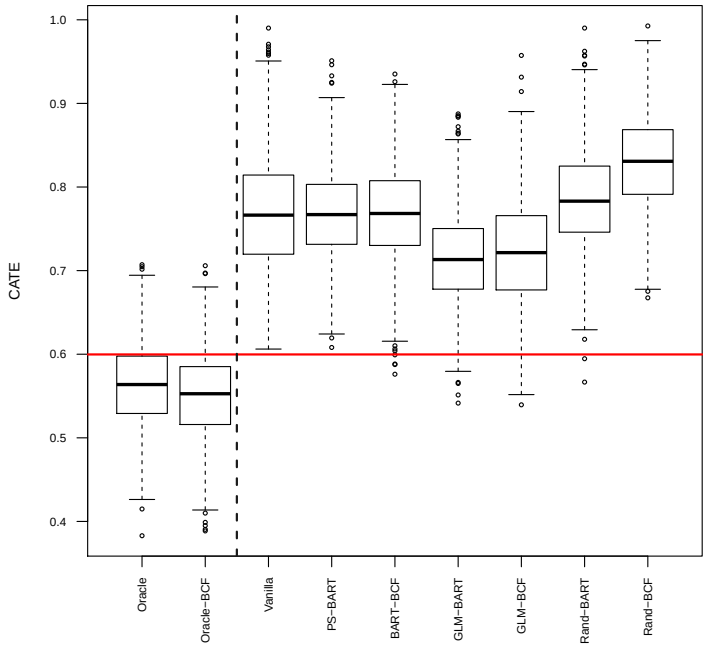

Figure 4.19: *Simulation Based on Real Data - Boxplots of CATE estimates over 1000 posterior draws for every model in one iteration. The models to the left of the dashed vertical line use the true propensity score as a covariate, thus those models are infeasible. The horizontal solid line is the true CATE for the sample.*

The convergence of the probit-BART used to estimate the propensity score was analyzed following Sparapani *et al.* (2019) instructions for performing Geweke convergence diagnostics (Geweke, 1992). For convenience, 10 individuals were sampled to have their ACF and trace plots analyzed. The ACF plot in Figure 4.20 shows that the individuals have almost no autocorrelation for the first 16 lags. The trace plot in Figure 4.21 seem to indicates that the response of the individuals traverse the sample space adequately. Furthemore, 92.2% of the Geweke statistics calculated for the whole sample are within the 95% range in Figure 4.22, which suggest that convergence might have been achieved, but can still be questioned.

For the remaining models, as instructed in Sparapani *et al.* (2019), the trace plots and the ACF plots of the σ draws for each model were analyzed. The plots can be found on Appendix C.1. In general, the trace plots seem to traverse the sample space adequately, but with some oscillations. The ACF presented some autocorrelation for all the models, so the convergence is questionable.

In general, the BART and the BCF models held similar results. The Vanilla model posterior CATE estimates apparently are impacted by the RIC phenomenon, so the model performed poorly. The Oracle models, as expected, had a good performance due to the inclusion of the true propensity score as a covariate. The PS and the GLM models performed slightly better than the Vanilla model, indicating that the inclusion of the estimated propensity score had a positive impact on the model,

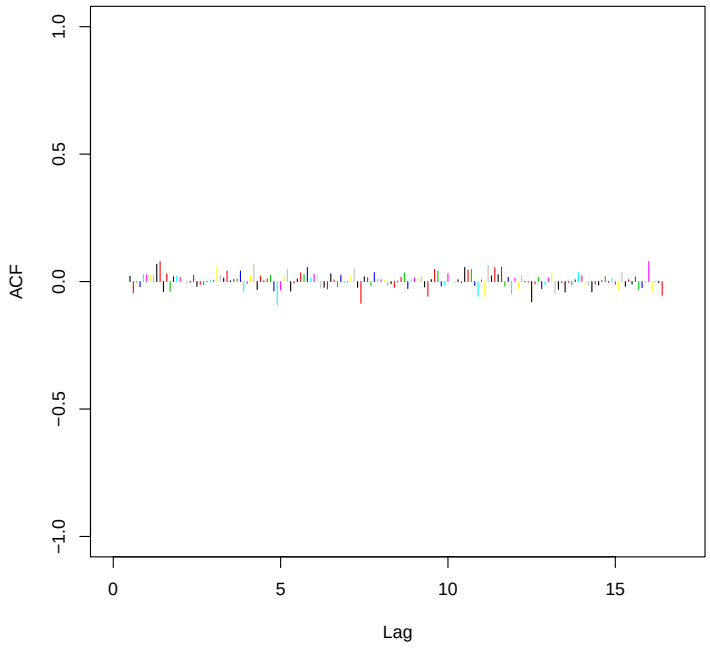

Figure 4.20: *Probit-BART (Simulation Based on Real Data) - ACF functions of 10 sampled individuals. Apparently, in this iteration the autocorrelation among the 1000 posterior draws is low in the sample.*

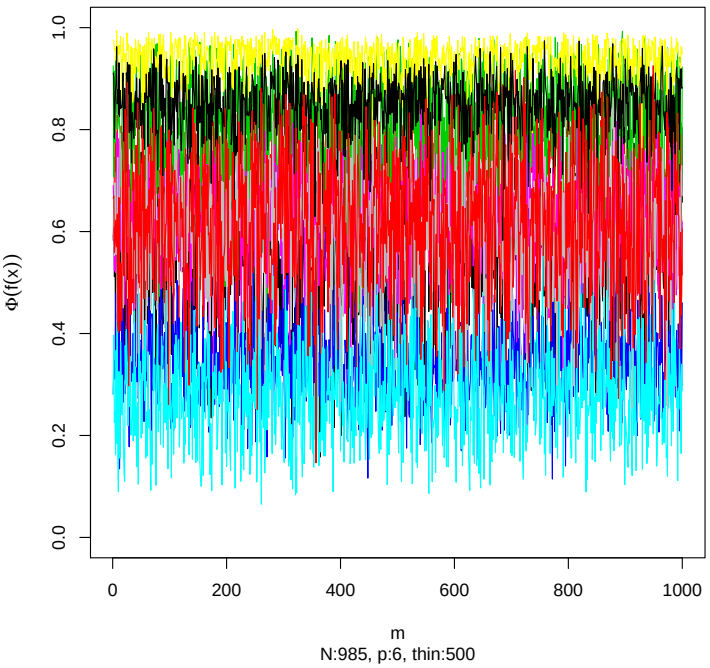

Figure 4.21: *Probit-BART (Simulation Based on Real Data) - Trace plots of 10 sampled individuals. Apparently, in this iteration the traces traverse the sample space adequately among the 1000 posterior draws.*

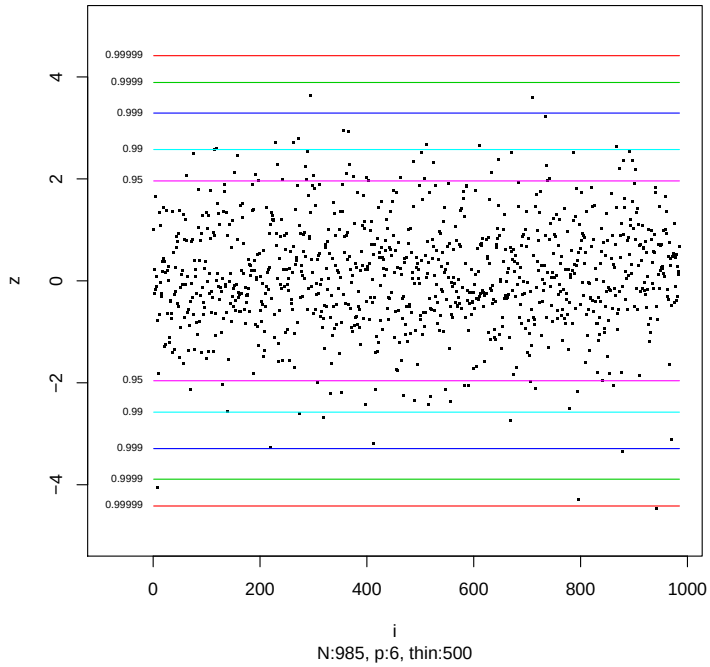

Figure 4.22: *Probit-BART (Simulation Based on Real Data) - Geweke statistics for each individual of the sample. Convergence is questionable in this iteration since 92.2% of the sample observations are within the 95% range.*

but the uncertainty associated with the estimation of propensity score contributed negatively on the CATE estimates. The Rand models held the worst results in the simulation due to the inclusion of an irrelevant variable and the lack of a propensity score estimate among its covariates.

As seen in the ICE Plots from the Appendix C.2, among the BART models, the inclusion of the true propensity score as a covariate greatly reduces the uncertainty over the individual treatment effects, allowing the visualization of different groups of individuals in the Oracle model and eliminating most of spurious effects from the other covariates. The inclusion of the estimated propensity score in the PS-BART and the GLM-BART seemed to slightly reduce the bias in the estimated CATE in relation to the Vanilla model. The Rand-BART ICE plots seemed similar to the Vanilla ICE plots. In general, the BCF models ICE plots seemed to indicate less uncertainty and correctly placing the individuals among the region where the true treatment effect for the groups was located. Also, since there was no interaction among the covariates, the c-ICE Plots and the d-ICE Plots individual curves should be in the same place as the PDP curve, but some spurious interactions were detected among the models, especially by the BCF models.

Across 1000 replications, the use of the propensity score as a covariate is corroborated, since the inclusion of true (estimated) propensity score greatly (slightly) improved the performance of the model, as seen in the CATE RMSE and ITE RMSE boxplots in Figures 4.23 and 4.24. The

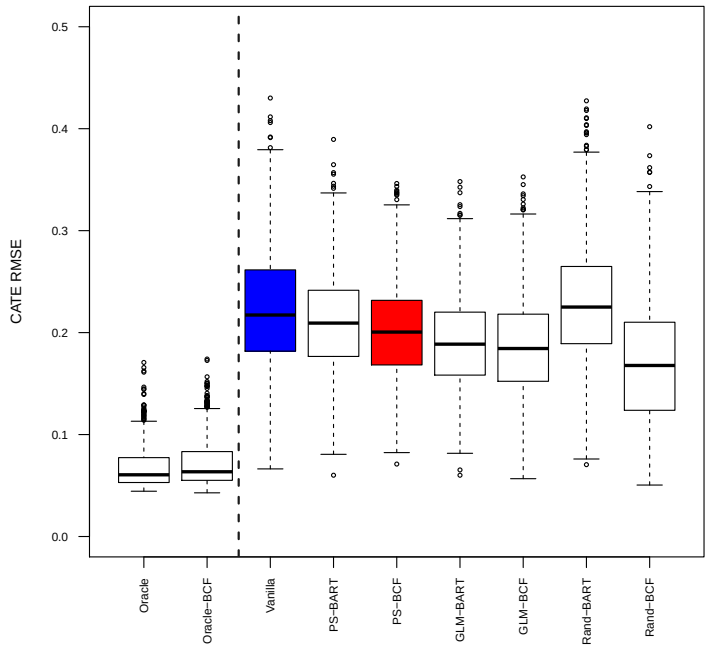

Figure 4.23: *Simulation Based on Real Data - Boxplots of the CATE RMSE for each model calculated over 1000 simulations. Vanilla (in blue) is the benchmark, while PS-BCF (in red) is the proposed model.*

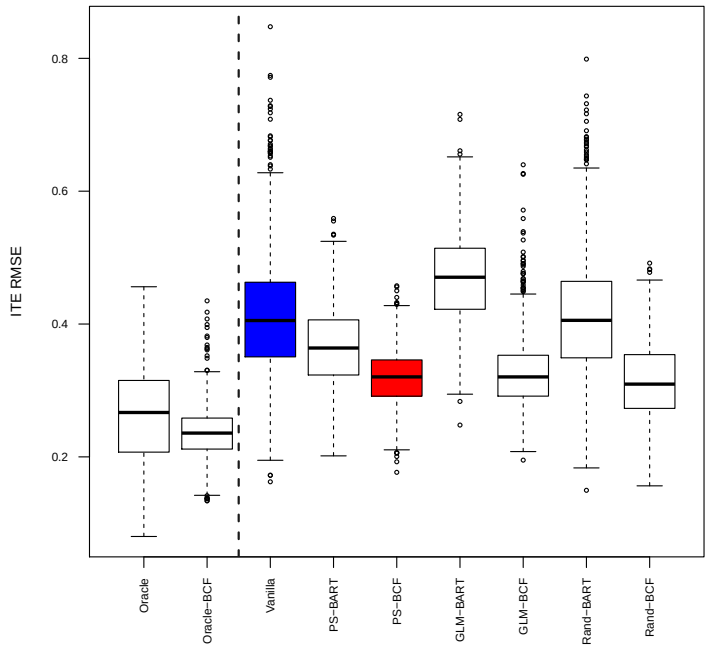

Figure 4.24: *Simulation Based on Real Data - Boxplots of the ITE RMSE for each model calculated over 1000 simulations. Vanilla (in blue) is the benchmark, while PS-BCF (in red) is the proposed model.*

Model	BART		BCF	
	CATE RMSE	ITE RMSE	CATE RMSE	ITE RMSE
Vanilla	0.223 (0.059)	0.411 (0.092)	- -	- -
Oracle	0.068 (0.020)	0.261 (0.075)	0.072 (0.023)	0.236 (0.039)
PS	0.210 (0.049)	0.364 (0.063)	0.201 (0.048)	0.319 (0.041)
GLM	0.191 (0.046)	0.469 (0.064)	0.186 (0.048)	0.328 (0.056)
Rand	0.229 (0.060)	0.412 (0.094)	0.171 (0.061)	0.314 (0.057)

Table 4.3: *Simulation Based on Real Data - Model assessment through the means of CATE RMSE, and ITE RMSE over replications. Standard deviation is given in parenthesis.*

means and the standard deviations for average RMSE of the CATE, as well as the ITE RMSE estimates for these models over the replications can be found on Table 4.3.

The BART and the BCF models held similar results across simulations, but for the ITE RMSE, the BCF model seems to have a better performance under the uncertainty of the estimation of the propensity score, despite having a higher variance across the the replications.

4.5.2 Sparse Data Example

Under the sparse setting, the following models were analyzed:

- Vanilla (Vanilla-DART): Y_i estimated by a BART (DART) model using $\tilde{x}_i$ as covariates;
- Oracle (Oracle-DART): Y_i estimated by a BART (DART) model using $\tilde{x}_i$ and $\pi(\tilde{x}_i)$, the true value of the propensity score, as covariates;
- PS-BART (PS-DART): Y_i estimated by a BART (DART) model using $\tilde{x}_i$ and $\hat{\pi}(\tilde{x}_i)$, estimated by the posterior mean of the probit-BART (probit-DART), as covariates;
- GLM-BART (GLM-DART): Y_i estimated by a BART (DART) model using $\tilde{x}_i$ and $\hat{\pi}(\tilde{x}_i)$, estimated by GLM, as covariates;
- Rand-BART (Rand-DART): Y_i estimated by a BART (DART) model using $\tilde{x}_i$ and $\hat{\pi}(\tilde{x}_i)$, given by a random Uniform distribution, as covariates;

Those simulations were replicated 1000 times in order to assess the results. In each model the first 50000 draws from the MCMC were treated as burn in, while the posterior draws had size 1000. The thinning used for the probit-BART was 500, while the other models had a thinning of 100.

In order to acknowledge the propensity score role in the model, a method of variable selection was performed. Selected variables were those whose presented $PIP > 0.5$. To assess the performance of the variable selection, following Linero (2018) and Bleich *et al.* (2014), Precision, Recall, and F_1 were used. These measures are defined by,

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN}, \qquad F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall},$$

where TP is True Positive, FP is False Positive, and FN is False Negative.

Friedman Function under Sparsity

The simulation adapted from Friedman (1991) example was generated as it follows,

$$Y_i = 10\sin(\pi x_{i1} x_{i2}) + 20(x_{i3} - 0.5)^2 + 10x_{i4} + 5x_{i5} + \mu_i + Z_i\alpha_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

$$x_{i1}, x_{i2}, ..., x_{i98} \sim \mathcal{U}(0, 1),$$

$$\mu_i = \mathbf{1}(x_{i1} < x_{i2}) - \mathbf{1}(x_{i1} \geq x_{i2}),$$

$$P(Z_i = 1 \mid x_{i1}, x_{i2}) = \Phi(\mu_i),$$

$$\alpha_i = 0.5 * \mathbf{1}(x_{i3} > 1/4) + 0.25 * \mathbf{1}(x_{i3} > 2/4) + 0.25 * \mathbf{1}(x_{i3} > 3/4),$$

$$\sigma = \frac{\theta^{(n)} - \theta^{(1)}}{8}, \qquad \theta_i = \mu_i + \alpha_i \Phi\left(\mu_i\right),$$

where $\theta^{(n)} = \max\left(\theta_1, ..., \theta_n\right)$ and $\theta^{(1)} = \min\left(\theta_1, ..., \theta_n\right)$.

The boxplots from Figure 4.25 are composed by posterior CATE for each model for one replication of the simulation. The horizontal line is the real CATE for this specific iteration. In general, the BART and DART models held very different results.

The Vanilla model posterior CATE seem to be biased by the RIC phenomenon, so the model performed poorly. The Oracle models, as expected, had a good performance with almost no bias due to the inclusion of the true propensity score as a covariate, but it should be noted that the inclusion of the Dirichlet hyperprior reduced the bias due to the removal of irrelevant covariates.

The PS models held the best performance among the analyzed models. The inclusion of the

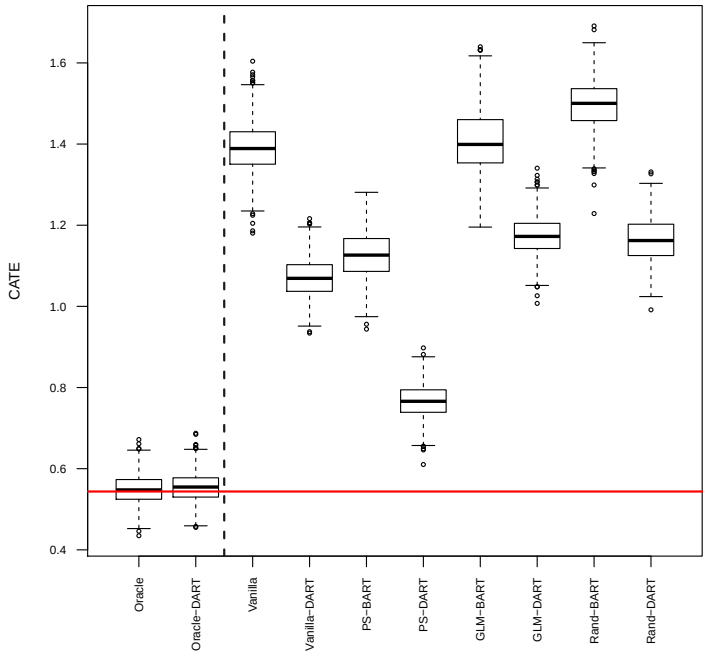

Figure 4.25: *Friedman Function under Sparsity - Boxplots of CATE estimates over 1000 posterior draws for every model in one iteration. The models to the left of the dashed vertical line use the true propensity score as a covariate, thus those models are infeasible. The horizontal solid line is the true CATE for the sample.*

propensity score as a covariate in the PS-BART model reduced the bias in the CATE estimates, but the inclusion of many irrelevant covariates still had a negative impact in the model. However, the PS-DART, due to the inclusion of the Dirichlet hyperprior, along with the inclusion of a propensity score estimated by the probit-DART model, greatly reduced the bias in the CATE estimates.

The GLM models performed poorly, indicating that the uncertainty associated with the estimation of propensity score contributed negatively on the CATE estimates. This effect might have happened due to the fact that the GLM do not perform variable selection or any kind of regularization on its default estimation.

The Rand models held the worst results in the simulation due to the inclusion of one more irrelevant covariate and the lack of a propensity score estimate among its covariates.

Again, the convergence of the probit-BART used to estimate the propensity score was analyzed following Sparapani *et al.* (2019) instructions for performing Geweke convergence diagnostics (Geweke, 1992). The propensity score used in the PS-BART was estimated by the probit-BART, while the propensity score used in the PS-DART was estimated by the probit-DART.

In regard to the probit-BART, it seems that the model did not achieved convergence. For convenience, 10 individuals were sampled to have their ACF and trace plots analyzed. The ACF plot

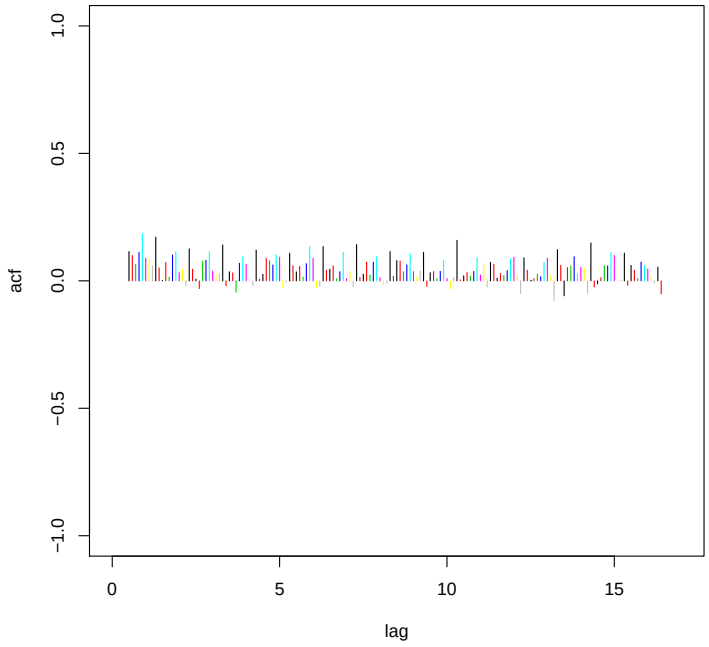

Figure 4.26: *Probit-BART (Friedman Function under Sparsity) - ACF functions of 10 sampled individuals. Apparently, in this iteration there is some autocorrelation among the 1000 posterior draws in the sample.*

in Figure 4.26 shows that the individuals have shown some autocorrelation for the first 16 lags. The trace plot in Figure 4.27 seem to indicates that the response of the individuals traverse the sample space with oscillations. The main problem is that 55.3% of the Geweke statistics calculated for the whole sample are within the 95% range in Figure 4.28, which strongly suggests that convergence have not been achieved.

Despite the convergence results for the probit-BART, the probit-DART seems to indicate that the model convergence might have been achieved. For convenience, 10 individuals were sampled to have their ACF and trace plots analyzed. The ACF plot in Figure 4.26 shows that the individuals have shown low autocorrelation for the first 16 lags. The trace plot in Figure 4.27 seem to indicates that the response of the individuals traverse the sample space adequately. Furthermore, 91.4% of the Geweke statistics calculated for the whole sample are within the 95% range in Figure 4.28, which suggest that convergence might have been achieved, but it might still be questionable.

All in all, the lack of convergence by the probit-BART is probably due to the inclusion of irrelevant covariates in the model, since the probit-DART Dirichlet hyperprior allow the model to adapt to the data with ease. As stated in the beginning of this section, only the covariates x_1 and x_2 were used to generate the propensity score, but, as can be seem on Figure 4.32, the probit-BART tends to use all covariates among the trees due to the uniform hyperprior, while, as expected, the

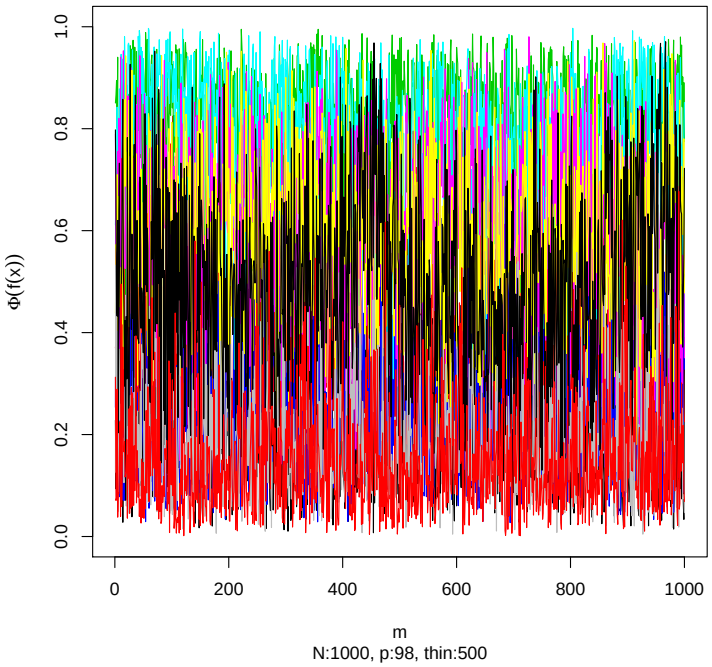

Figure 4.27: *Probit-BART (Friedman Function under Sparsity) - Trace plots of 10 sampled individuals. Apparently, in this iteration the traces traverse the sample space with oscillations among the 1000 posterior draws.*

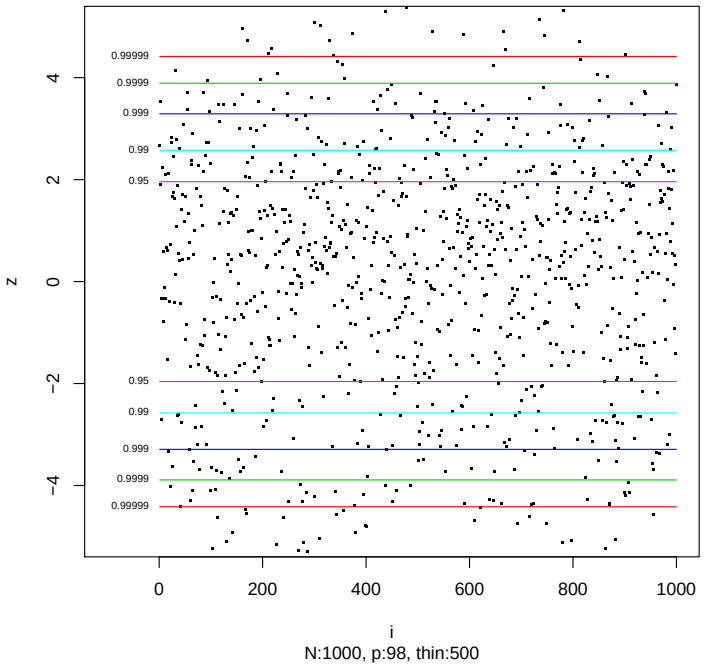

Figure 4.28: *Probit-BART (Friedman Function under Sparsity) - Geweke statistics for each individual of the sample. Convergence is questionable in this iteration since 55.3% of the sample observations are within the 95% range.*

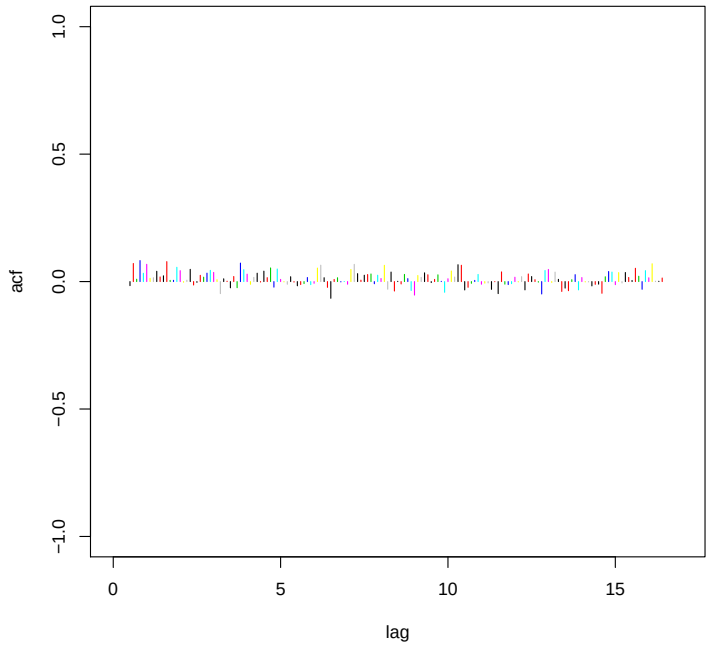

Figure 4.29: *Probit-DART (Friedman Function under Sparsity) - ACF functions of 10 sampled individuals. Apparently, in this iteration the autocorrelation among the 1000 posterior draws is low in the sample.*

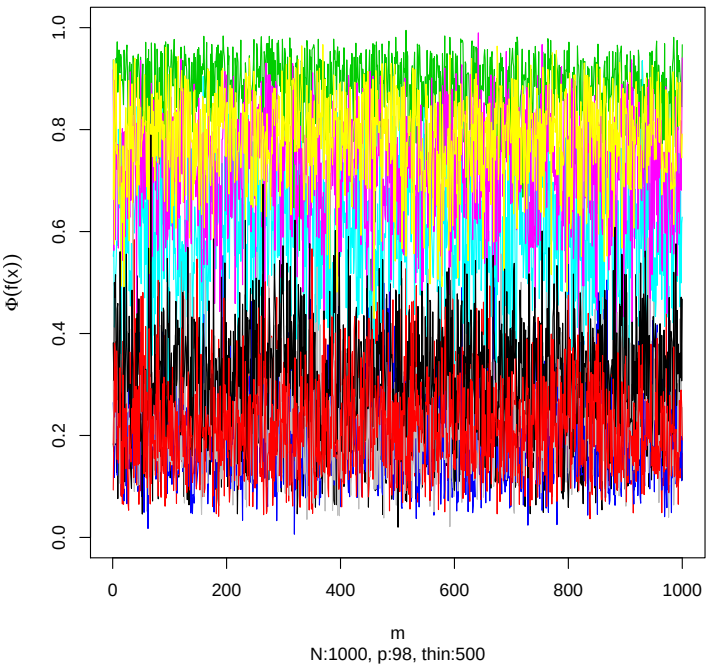

Figure 4.30: *Probit-DART (Friedman Function under Sparsity) - Trace plots of 10 sampled individuals. Apparently, in this iteration the traces traverse the sample space adequately among the 1000 posterior draws.*

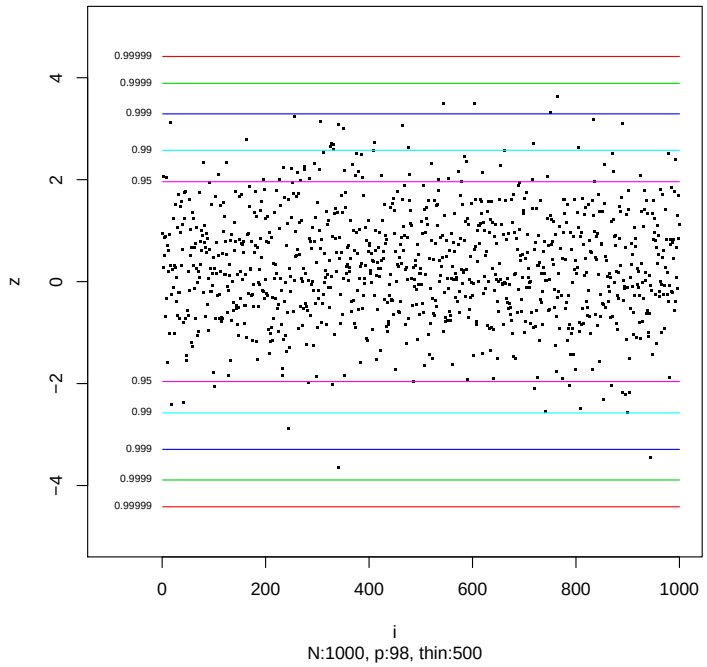

Figure 4.31: *Probit-DART (Friedman Function under Sparsity) - Geweke statistics for each individual of the sample. Convergence is questionable in this iteration since 91.4% of the sample observations are within the 95% range.*

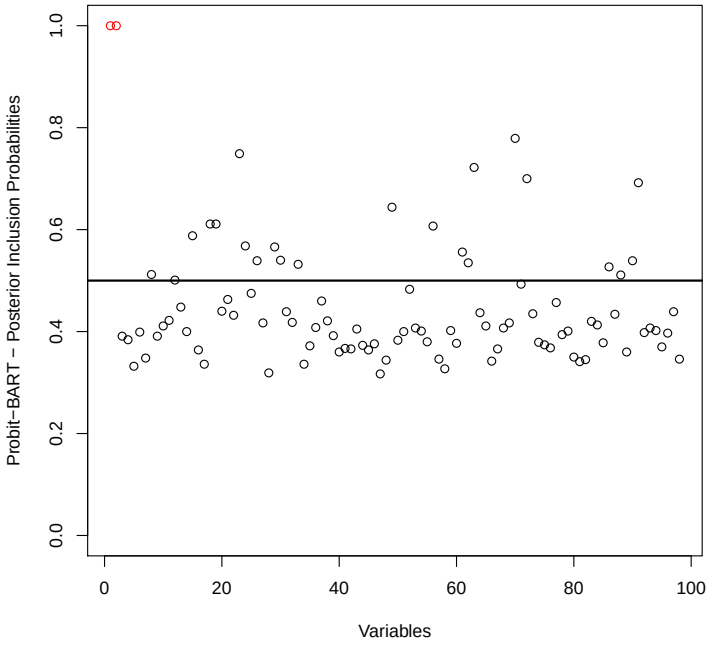

Figure 4.32: *Friedman Function under Sparsity - Posterior Inclusion Probability of Probit-BART model. In red: x_1, and x_2, respectively.*

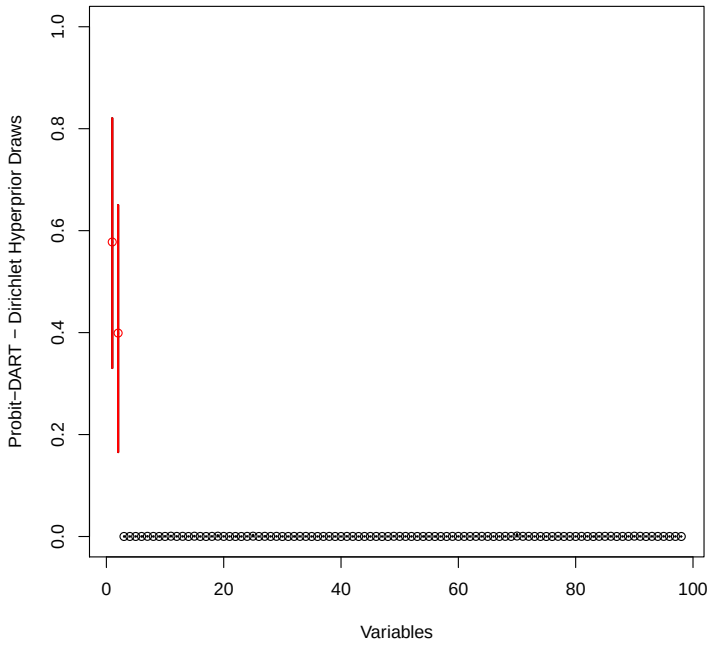

Figure 4.33: *Friedman Function under Sparsity - Posterior draws from Probit-DART Dirichlet hyperprior with 95% credible intervals. In red:* x_1*, and* x_2*, respectively.*

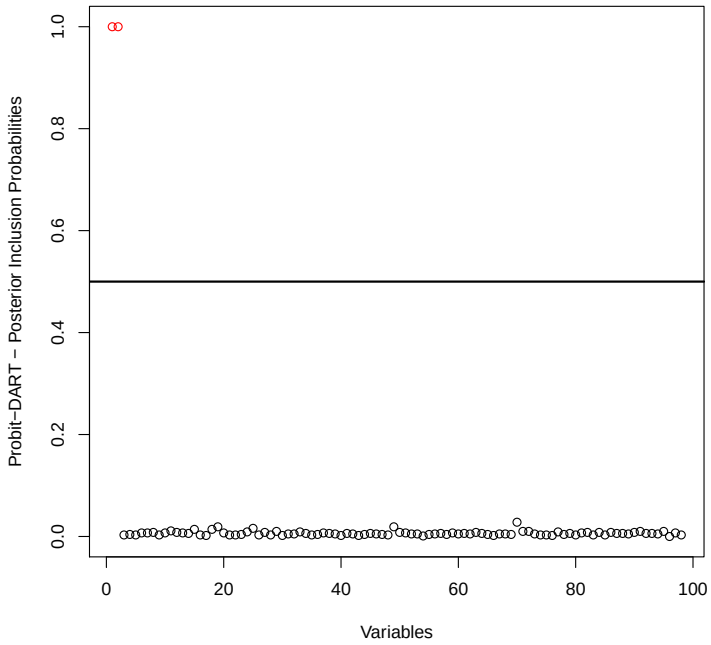

Figure 4.34: *Friedman Function under Sparsity - Posterior Inclusion Probability of Probit-DART model. In red:* x_1*, and* x_2*, respectively.*

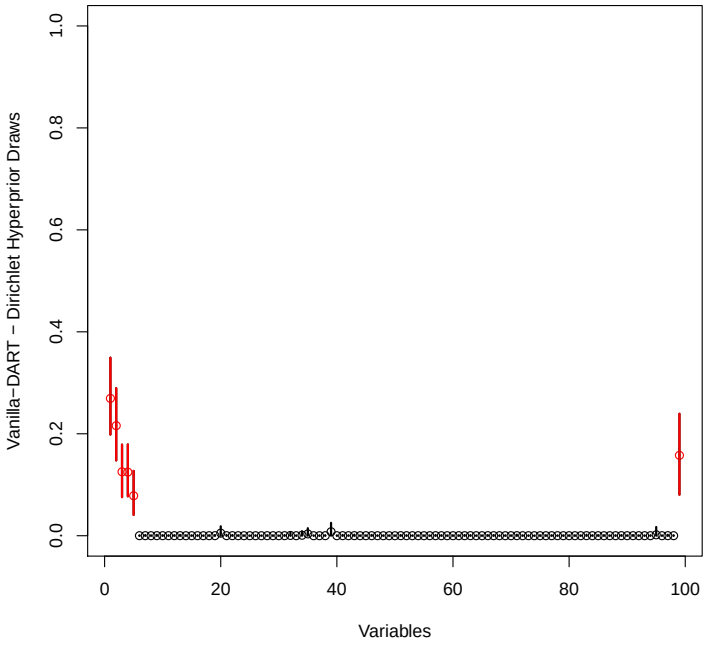

Figure 4.35: *Friedman Function under Sparsity - Posterior draws from Vanilla-DART Dirichlet hyperprior with 95% credible intervals. In red:* x_1, x_2, x_3, x_4, x_5, *and* z, *respectively.*

probit-DART correctly selected the covariates, as seem on Figures 4.33 and 4.34

Figures 4.35 and 4.36 represents the Dirichlet hyperprior draws for the Vanilla-DART and PS-DART models, while the Figures 4.37 and 4.38 represents the respective PIP estimations for those models. In the figures for this iteration, both DART models mostly select as relevant variables to the model only true data generating variables and the propensity score. The Vanilla-DART model presented three false positives in this iteration, but zero false negatives, while the PS-DART accurately selected only the relevant covariates, along with the propensity score. The variable selection graphs for the other models can be found on Appendix D.1, but for short, it seems that the DART models mostly selected only the relevant covariates, while the BART models are induced to selected all the covariates due to the uniform hyperprior.

As seen in the ICE Plots from the Appendix D.2, the BART models, except for the Oracle, seemed to fail at capturing both the CATE and the ITE among the models, as well as indicating spurious relations between the variables through the c-ICE and d-ICE plots. The model that held the best performance among the BART models was the PS-BART. Most of the feasible DART models have also shown some difficulty at capturing CATE and ITE, but the PS-DART had a surprisingly good performance, capturing most of the effects caused by variable x_3 and correctly placing the true CATE inside the 95% credible interval for the PDP for the five covariates analyzed.

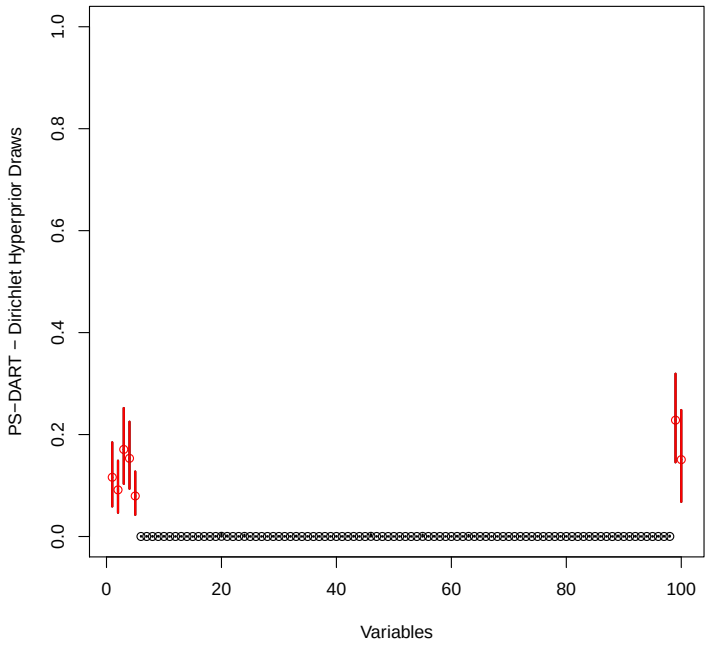

Figure 4.36: *Friedman Function under Sparsity - Posterior draws from PS-DART Dirichlet hyperprior with 95% credible intervals. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

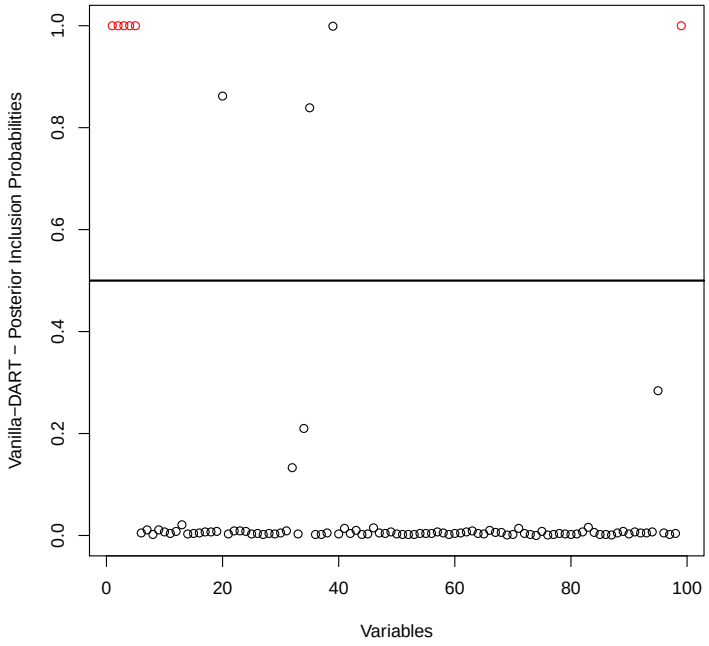

Figure 4.37: *Friedman Function under Sparsity - Posterior Inclusion Probability of Vanilla-DART model. In red:* x_1, x_2, x_3, x_4, x_5, *and* z, *respectively.*

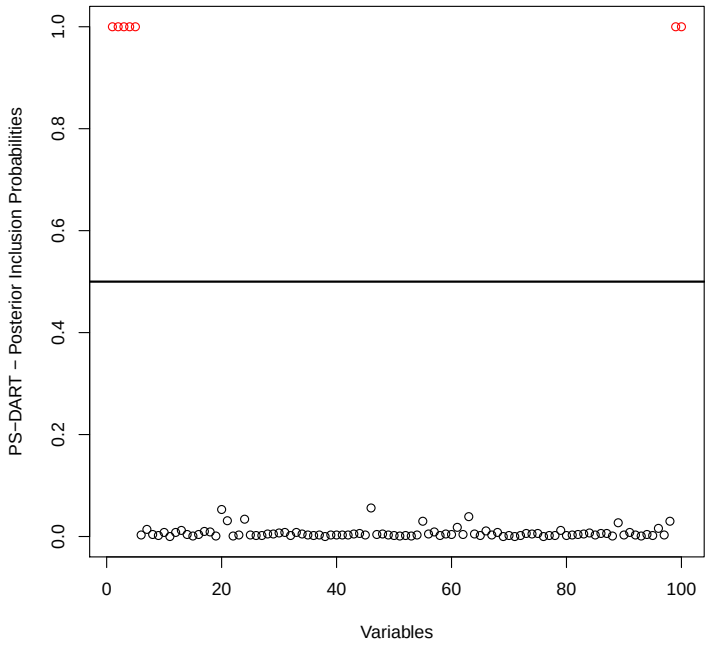

Figure 4.38: *Friedman Function under Sparsity - Posterior Inclusion Probability of PS-DART model. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$, *and* z, *respectively.*

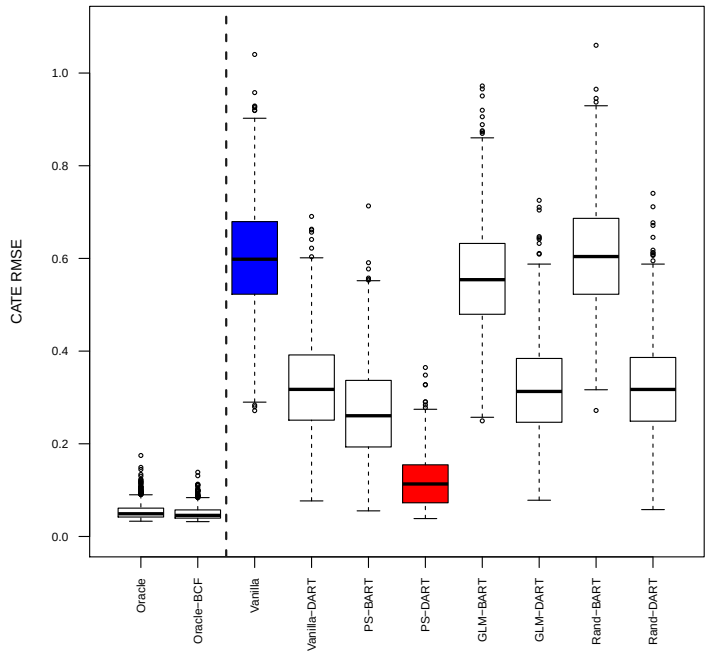

Figure 4.39: *Friedman Function under Sparsity - Boxplots of the CATE RMSE for each model calculated over 1000 simulations. Vanilla (in blue) is the benchmark, while PS-DART (in red) is the proposed model.*

	Model	CATE RMSE	ITE RMSE	Precision	Recall	F_1	PS-Usage
BART	Vanilla	0.603	1.223	0.061	1.000	0.115	-
		(0.117)	(0.192)	(0.001)	(0.000)	(0.001)	-
	Oracle	0.054	0.472	0.095	1.000	0.173	100.000
		(0.018)	(0.042)	(0.007)	(0.000)	(0.012)	(0.000)
	PS	0.268	0.785	0.072	1.000	0.133	100.000
		(0.104)	(0.157)	(0.002)	(0.000)	(0.003)	(0.000)
	GLM	0.559	1.158	0.071	1.000	0.132	99.675
		(0.115)	(0.180)	(0.001)	(0.000)	(0.002)	(0.021)
	Rand	0.606	1.229	0.060	1.000	0.114	-
		(0.118)	(0.192)	(0.001)	(0.000)	(0.002)	-
	Probit	-	-	0.097	1.000	0.177	-
		-	-	(0.014)	(0.000)	(0.023)	-
DART	Vanilla	0.323	1.007	0.861	1.000	0.919	-
		(0.105)	(0.233)	(0.139)	(0.000)	(0.086)	-
	Oracle	0.051	0.468	0.989	1.000	0.994	100.000
		(0.015)	(0.036)	(0.039)	(0.000)	(0.022)	(0.000)
	PS	0.119	0.677	0.770	1.000	0.861	100.000
		(0.055)	(0.130)	(0.156)	(0.000)	(0.103)	(0.000)
	GLM	0.320	1.000	0.877	0.905	0.884	35.304
		(0.102)	(0.224)	(0.126)	(0.067)	(0.072)	(0.440)
	Rand	0.322	1.003	0.862	1.000	0.920	-
		(0.103)	(0.227)	(0.137)	(0.000)	(0.084)	-
	Probit	-	-	0.985	1.000	0.991	-
		-	-	(0.071)	(0.000)	(0.043)	-

Table 4.4: *Friedman Function under Sparsity - Model assessment through the means of CATE RMSE, ITE RMSE, Precision, Recall, F_1, and usage of the propensity score over replications. Standard deviation is given in parenthesis.*

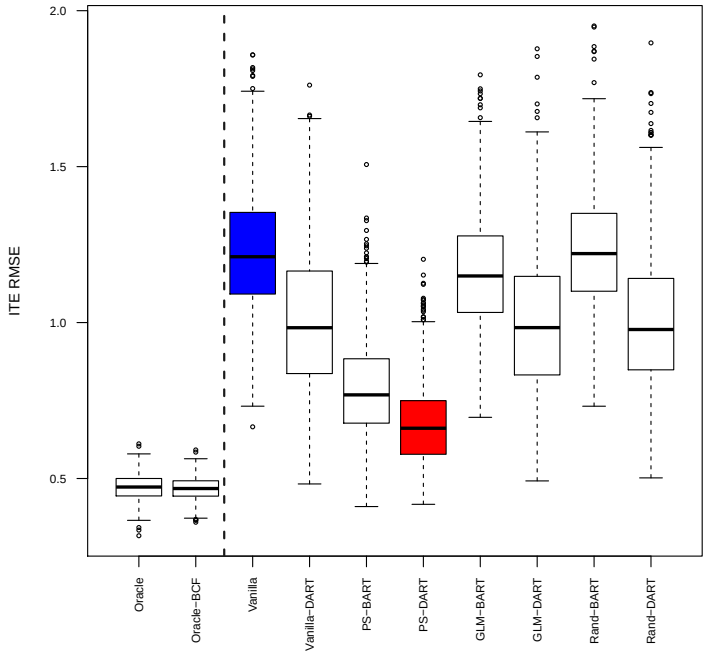

Figure 4.40: *Friedman Function under Sparsity - Boxplots of the ITE RMSE for each model calculated over 1000 simulations. Vanilla (in blue) is the benchmark, while PS-DART (in red) is the proposed model.*

The performance of the models over replications is evaluated at Table 4.4, along with Figures 4.39 and 4.40. Variables were selected via the $PIP > 0.5$ criteria. For Precision, Recall and F_1, the value 1.0 indicates perfect adjustment. PS-Usage indicates the proportion of times over the replications that the propensity score estimation was selected as a relevant variable in the model. All measurements are given by the mean over the replications, with standard deviation in parentheses.

4.5.3 Simulations Assessment

The simulations in Sections 4.5.1 and 4.5.2 were performed in order to assess the inclusion of the propensity score as a covariate and advocate for tools that can be used on treatment effect analysis.

In all examples the inclusion of the true or the estimated propensity score, if correctly estimated, resulted in a decrease of the impact that the RIC phenomenon had over the model. As expected, the models which had the true propensity score have shown almost no bias. The two models used to estimated the propensity score, namely the probit-BART and the GLM, have shown similar performance in the simulation based on real data, but on the sparsity examples the BART based models had superior results. That may be due to the fact that the GLM was including all the variables in the model, while the BART and the DART models can naturally incorporate interac-

tions between covariates, and even perform accurate variable selection in the case of DART. The simulations were generated in a simple setting, allowing both models to adjust relatively well, but in real datasets there might be unusual interactions between the covariates, as well as irrelevant variables, which is a scenario that models derived from BART, such as DART and BCF, can adapt with ease.

The flexibility of the ICE Plot allows it to be used under many different scenarios, but in the treatment effect setting it brings up three main advantages: allows the visualization of variables that do not impact in the treatment effect; show possible candidates of relevant variables for different individuals; and grants a way that may help in the identification of groups whose individuals may be affected in different ways by the chosen treatment.

Under the sparsity setting, it can be seen that the DART models variable selection performed well. Moreover, based on the results, the PIP from the DART model can be considered an important tool upon the definition of which variables to include in the propensity score estimation.

4.6 Real Data Analysis

In order to corroborate the use of the tools introduced in this Chapter, the 1987 National Medical Expenditure Survey (NMES) dataset from Johnson *et al.* (2003) is analyzed. The reason for this dataset to be chosen is that since it has already been analyzed by Hahn *et al.* (2018b), then by using the methodology introduced here it should be possible to identify different groups of individuals and compare these groups with the ones identified by these authors. One important note is that it is not possible to ensure *strong ignorability* on observational studies, so the treatment effects might be biased due to some unmeasured confounders.

Covariates with multiple factors were transformed in multiple *dummy* covariates since the actual implementation of the BCF is not able to deal with factors. Some of these covariates are ordered factors, but were not transformed in numeric covariates since it is easier to analyze the effect of each level separately.

The following covariates were selected to be used in the model:

- *packyears*: Number of years of pack-a-day smoking $\left(\frac{\text{number of cigarretes per day}}{20} \times \text{number of years smoked}\right)$;
- *LASTAGE*: Age of the individual at the time of the survey;
- *AGESMOKE*: Age that the individual started smoking;
- *MALE*: 1 if the individual is male, 0 if female;

- *RACE3.1*: 1 if the individual is white, 0 otherwise;
- *RACE3.2*: 1 if the individual is black, 0 otherwise;
- *RACE3.3*: 1 if the individual is not black or white, 0 otherwise;
- *marital.1*: 1 if the individual is married, 0 otherwise;
- *marital.2*: 1 if the individual is widowed, 0 otherwise;
- *marital.3*: 1 if the individual is divorced, 0 otherwise;
- *marital.4*: 1 if the individual is separated, 0 otherwise;
- *marital.5*: 1 if the individual never married, 0 otherwise;
- *educate.1*: 1 if the individual has college graduate, 0 otherwise;
- *educate.2*: 1 if the individual has some college, 0 otherwise;
- *educate.3*: 1 if the individual has high school graduate, 0 otherwise;
- *educate.4*: 1 if the individual has some other level of education, 0 otherwise;
- *SREGION.1*: 1 if the census region is Northeast, 0 otherwise;
- *SREGION.2*: 1 if the census region is Midwest, 0 otherwise;
- *SREGION.3*: 1 if the census region is South, 0 otherwise;
- *SREGION.4*: 1 if the census region is West, 0 otherwise;
- *POVSTALB.1*: 1 if the individual is poor, 0 otherwise;
- *POVSTALB.2*: 1 if the individual is near poor, 0 otherwise;
- *POVSTALB.3*: 1 if the individual has low income, 0 otherwise;
- *POVSTALB.4*: 1 if the individual has middle income, 0 otherwise;
- *POVSTALB.5*: 1 if the individual has high income, 0 otherwise;
- *beltuse.1*: 1 if the individual rarely uses the seatbelt, 0 otherwise;
- *beltuse.1*: 1 if the individual sometimes uses the seatbelt, 0 otherwise;
- *beltuse.1*: 1 if the individual always/almost always uses the seatbelt, 0 otherwise;

- *yearsince*: Years since the individual quit smoking;
- *TOTALEXP*: Annual medical expenditures.

Following Imai and van Dyk (2004) and Hahn *et al.* (2018b) approaches, the treatment Z was defined as 1 for individuals with *packyears* greater than 17, and 0 otherwise. Only individuals with positive medical expenditures and had LASTAGE greater or equal to 28 were considered in the sample. Individuals that had a *missing value* for any of the covariates presented previously were discarded from the sample, resulting in a sample with 6874 individuals. The vector of all covariates, except *packyears* and *TOTALEXP* for the ith individual os the sample, $i = (1, ..., 6874)$, is denoted as $\tilde{x}_i$. The natural logarithm of *TOTALEXP* for the ith individual is denoted as Y_i, which means that the treatment effect for Y_i will have a multiplicative effect on *TOTALEXP*.

The following models were constructed:

- Probit-BART: $\hat{\pi}(\tilde{x}_i)$ estimated by a probit-BART model using $\tilde{x}_i$ as covariates;
- DART: Y_i estimated by a DART model using $\tilde{x}_i$ as covariates;
- BCF: Y_i estimated by a BCF model using $\tilde{x}_i$ and $\hat{\pi}(\tilde{x}_i)$, the estimated propensity score, as covariates;

The models had a burn-in of 15000 draws, thinning of 100 draws and a posterior size of 2000 draws. The only exception is the probit-BART, which used thinning of 5000 draws due to the difficulty of the model to achieve convergence, even so, like it can be seem on Figures 4.41, 4.42 and 4.43, convergence is still questionable.

In regard to the convergence of the DART and the BCF models, the trace plots of both models seemed to traverse the sample space adequately as seem on Figures 4.44 and 4.45. The ACF plots of both models indicate that there is almost no autocorrelation between the σ draws, as seem on Figures 4.46 and 4.47.

The Dirichlet hyperprior draws and the PIP for the DART model seemed inconclusive. These graphics can been found on Appendix E.1.

The CATE boxplots showed similar results for both models, but as it can be noted, the DART models apparently carries more uncertainty among the estimations of individual treatment effects in relation to the BCF model. The CATE boxplots can be seem on Figure 4.48, while the estimated ITE and the respective 95% credible intervals for each model can be seem on Figures 4.49 and 4.50.

Since the BCF model seem to have less uncertainty and shows some individuals above the 95% credible interval for the estimated ITE, the ICE plots for this model are analyzed.

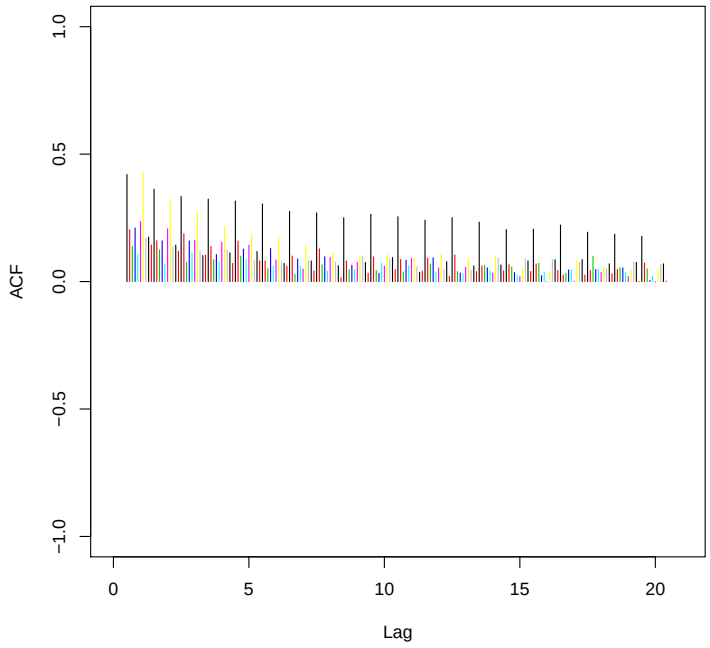

Figure 4.41: *Probit-BART (Real Data Analysis) - ACF functions of 10 sampled individuals. Apparently, in this iteration there is some autocorrelation among the 2000 posterior draws in the sample.*

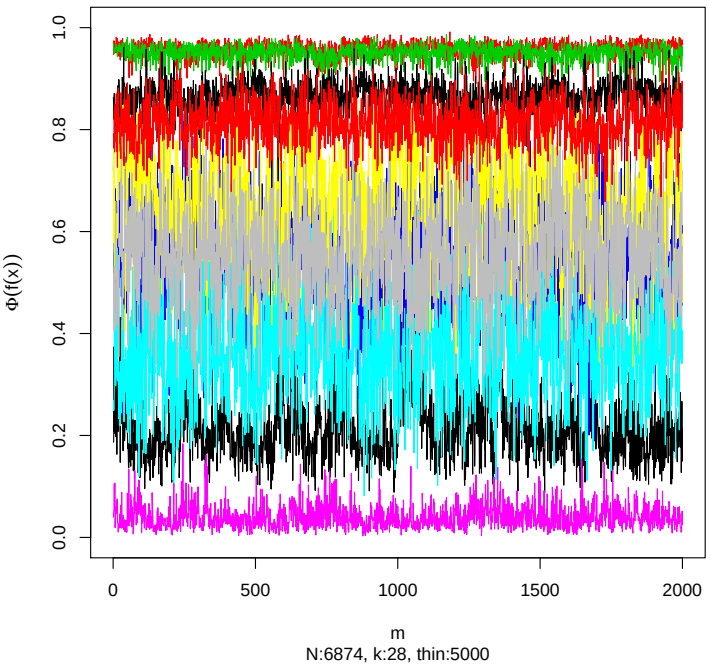

Figure 4.42: *Probit-BART (Real Data Analysis) - Trace plots of 10 sampled individuals. Apparently, in this iteration the traces traverse the sample space with oscillations among the 2000 posterior draws.*

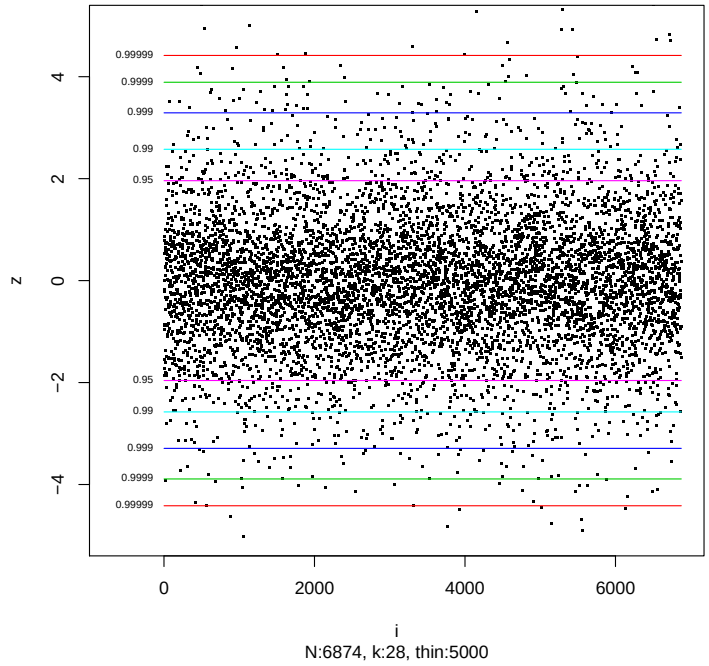

Figure 4.43: *Probit-BART (Real Data Analysis) - Geweke statistics for each individual of the sample. Convergence is questionable in this iteration since 87.5% of the sample observations are within the 95% range.*

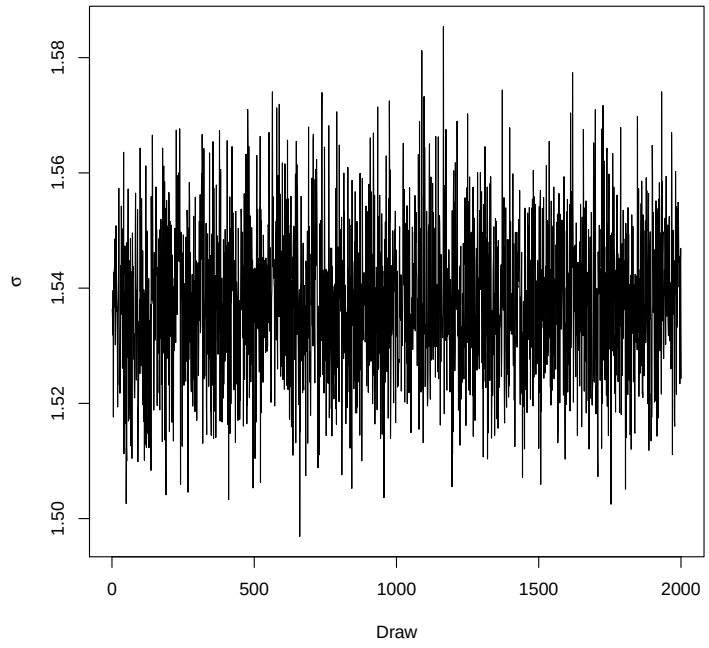

Figure 4.44: *DART model (Real Data Analysis) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

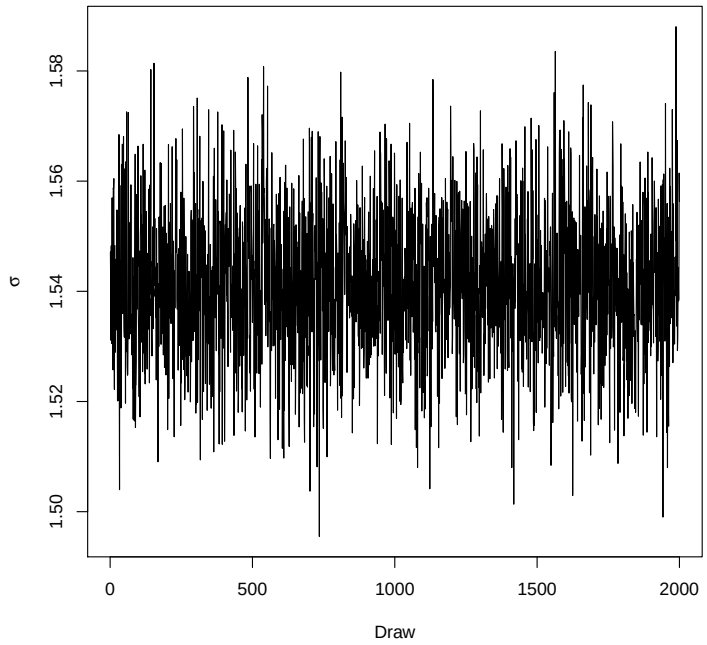

Figure 4.45: *BCF model (Real Data Analysis) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

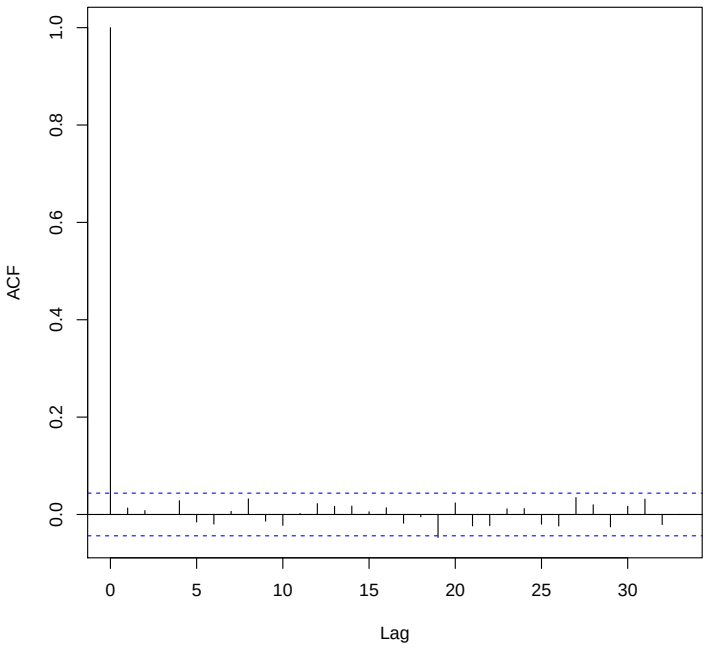

Figure 4.46: *DART model (Real Data Analysis) - ACF function for the σ draws. Apparently there is almost no autocorrelation among the draws.*

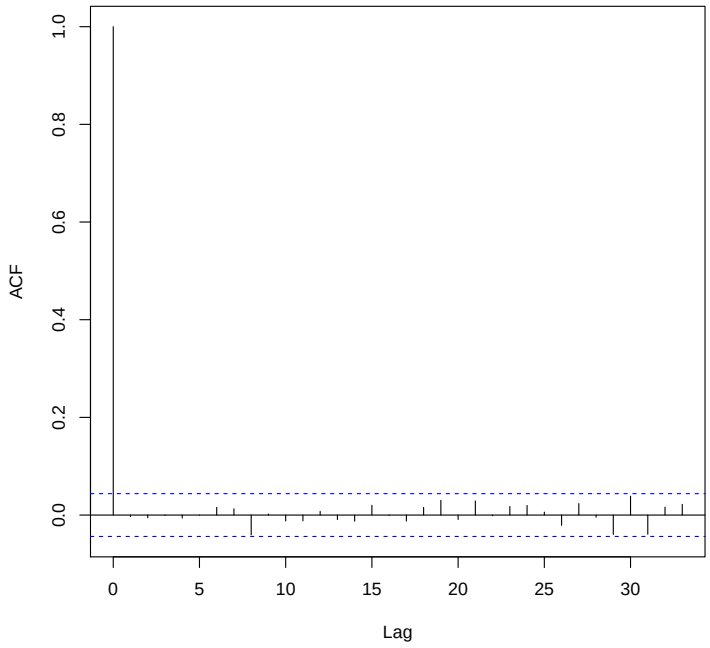

Figure 4.47: *BCF model (Real Data Analysis) - ACF function for the σ draws. Apparently there is almost no autocorrelation among the draws.*

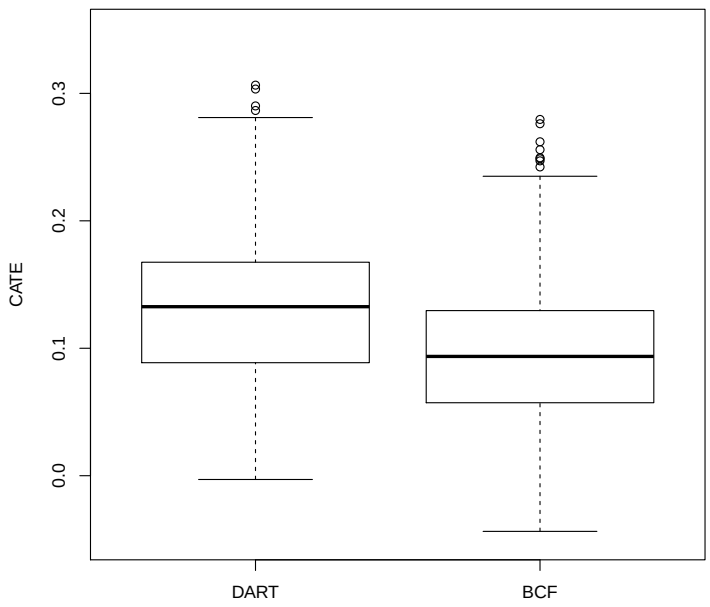

Figure 4.48: *Real Data Analysis - Boxplots of the estimated CATE over* 2000 *posterior draws for the DART model and for the BCF model.*

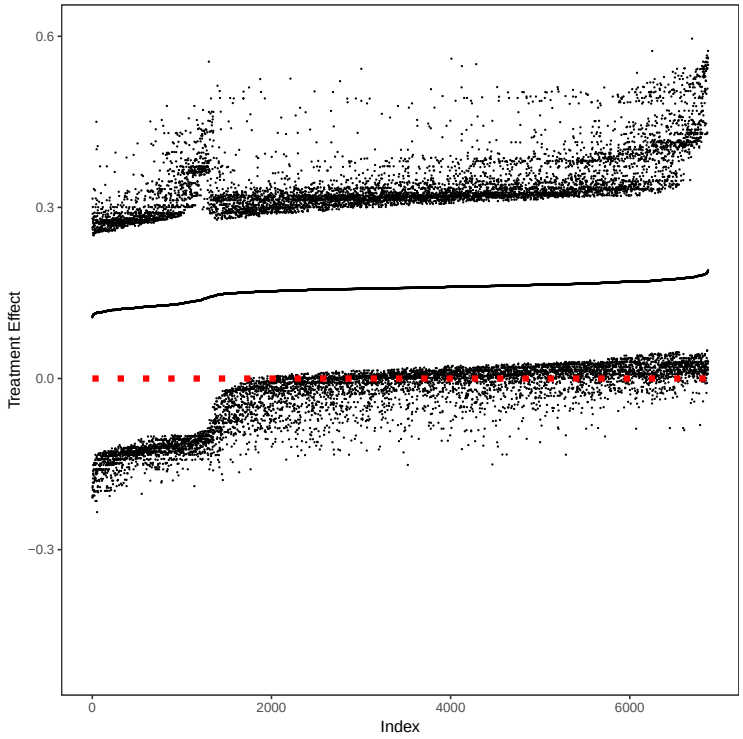

Figure 4.49: *DART model (Real Data Analysis) - Ordered posterior mean of the ITE with 95% credible intervals. The red dotted line indicates the point where the treatment effect axis is zero.*

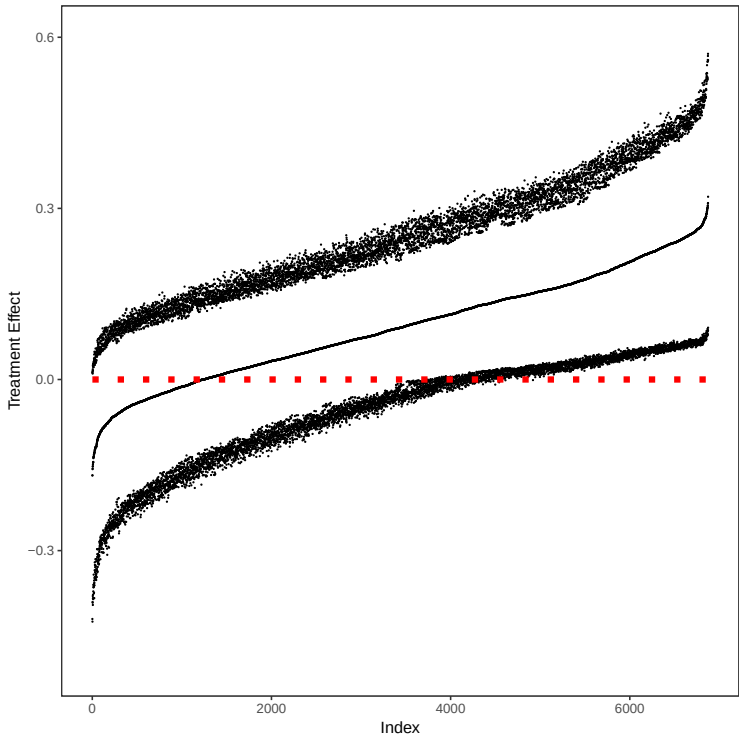

Figure 4.50: *BCF model (Real Data Analysis) - Ordered posterior mean of the ITE with 95% credible intervals. The red dotted line indicates the point where the treatment effect axis is zero.*

The ICE plots every variable used in the model is shown on Appendix E.2. As it can be seem in the ICE and c-ICE plots, most of the covariates seemed to be unrelated to the treatment effect. Only four covariates presented some change: $LASTAGE$, $MALE$, $marital.1$, and $POVSTALB.1$.

Among those covariates, the main interest lies in $LASTAGE$. By performing an initial analysis it seems that the individuals from the sample can be divided in groups. The points 45.5, 53.5, 64.5 and 73.5 were selected possible candidates to partition the data, as seem on Figure 4.51. Since the individuals with $LASTAGE$ lower than 46 presented considerable high levels of treatment effect in comparison to the remaining values in the ICE plot, the individuals with those traits were identified. As it can be seem on Figure 4.52 these individuals, in general, had positive treatment effect estimates, many of those did not included 0 in the 95% credible intervals.

Then groups of individuals were formed among the covariate $LASTAGE$, but since the c-ICE and the d-ICE plots suggest that the estimated treatment effect might be related to other covariates, the remaining three covariates were analysed in junction to the $LASTAGE$ c-ICE plot.

The effect observed on $POVSTALB.1$ seem to not be related with $LASTAGE$ since, as seem on Figure 4.53, the individuals with these traits have not shown any pattern over the c-ICE plot.

Different from the previous covariate, $marital.1$ seems to show some ICE curves concentrated around the PDP curve when the individual is married, and dispersing otherwise, as seem on Figure 4.54.

In regard to the covariate $MALE$, the ICE curves have practically been divided by the PDP curve. In most cases, if $MALE = 0$, then the ICE curve for this specific individual will be placed above the the PDP curve, and below if $MALE = 1$, as shown on Figure 4.55.

Clearly the last two covariates have shown signs of the existence of groups with different treatment effects. By analyzing $MALE$ and $marital.1$ in conjunction with $LASTAGE$ c-ICE plot four candidates for groups show up. Each group reacts to the treatment effect with different intensity, as seem on Figure 4.56.

The main advantage of this approach is that, different from Hahn *et al.* (2018b), which fitted the model of a fitted model, we simply use the trees draws to perform a kind of sensitivity analysis, relying only on the predictive performance of a single model. By using this methodology we were able to identify many groups of individuals that seem to have different treatment effects depending on the covariates observed for the individual. Furthermore, since the results from this analysis are similar to those found by Hahn *et al.* (2018b), the ICE plots have shown to be an interesting and useful tool for treatment effect analysis.

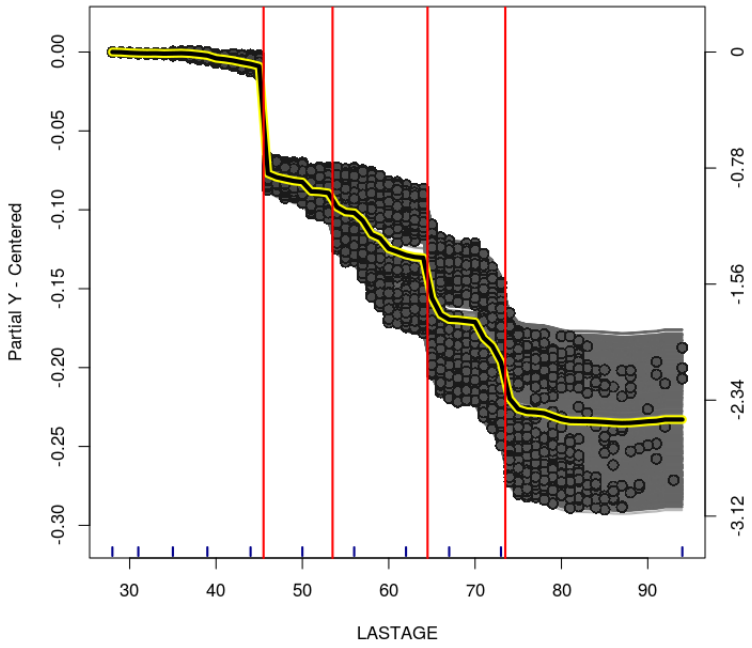

Figure 4.51: *Real Data Analysis - Centered ICE Plot for LASTAGE. The red vertical lines cut the LASTAGE axis at* 45.5, 53.5, 64.5 *and* 73.5, *respectively.*

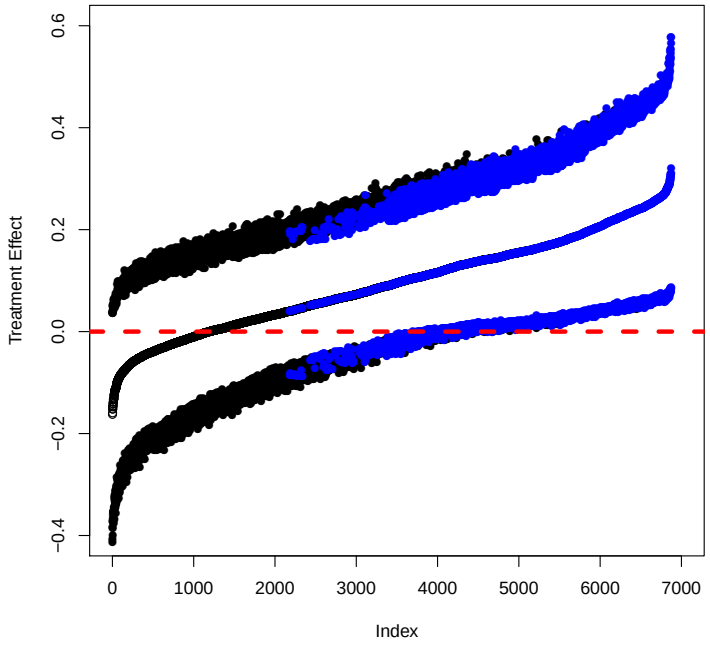

Figure 4.52: *Real Data Analysis - Ordered posterior mean of the ITE with 95% credible intervals. The red dotted line indicates the point where the treatment effect axis is zero. The blue observations denote individuals whose covariate LASTAGE is lower than 46.*

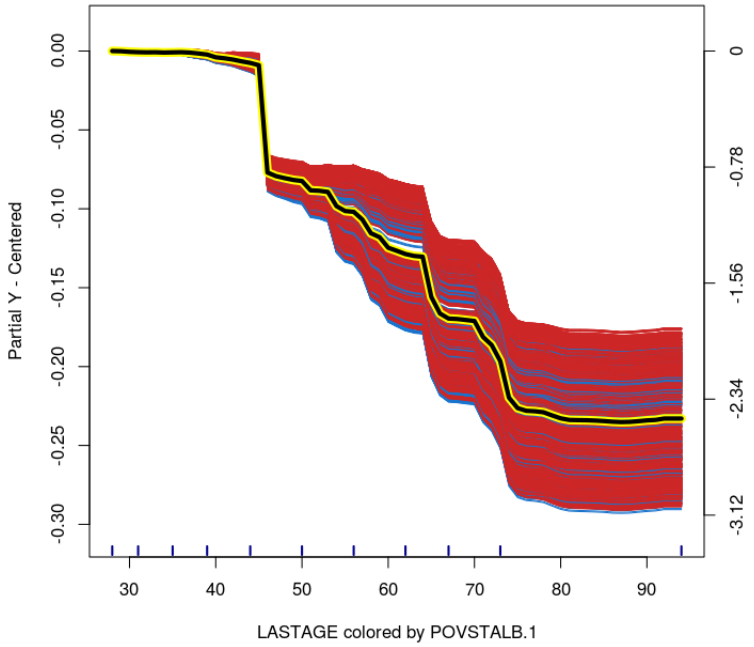

Figure 4.53: *Real Data Analysis - Centered ICE Plot for LASTAGE. If POVSTALB.1=0, the ICE curve is red, and blue otherwise.*

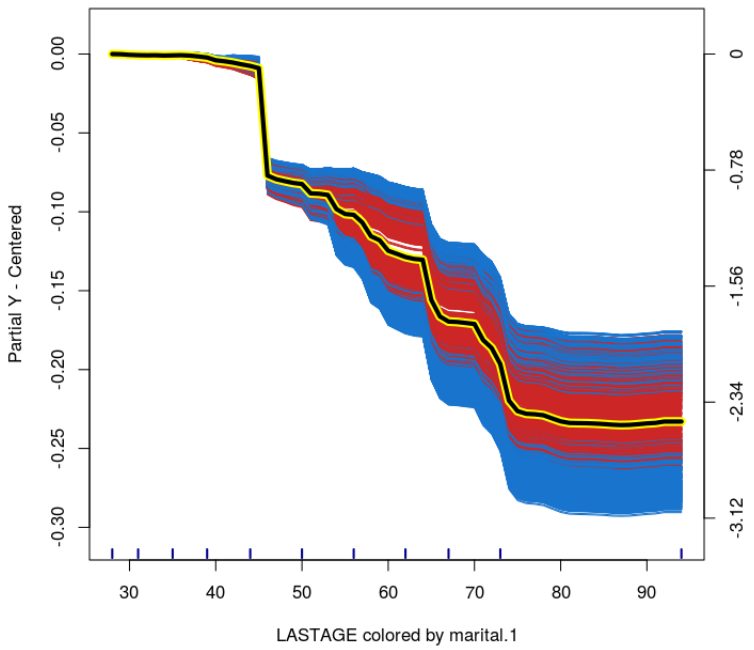

Figure 4.54: *Real Data Analysis - Centered ICE Plot for LASTAGE. If marital.1=0 and MALE=0, the ICE curve is red, and blue otherwise.*

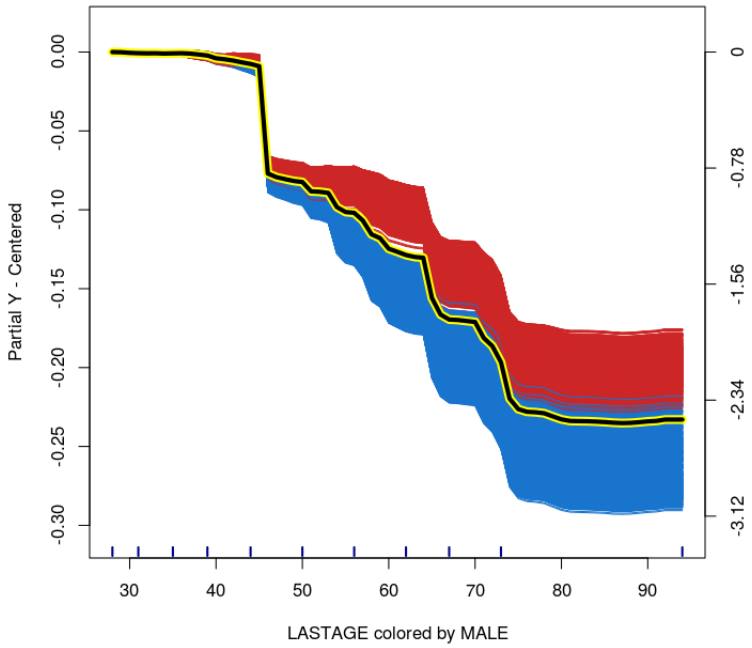

Figure 4.55: *Real Data Analysis - Centered ICE Plot for LASTAGE. If MALE=0, the ICE curve is red, and blue otherwise.*

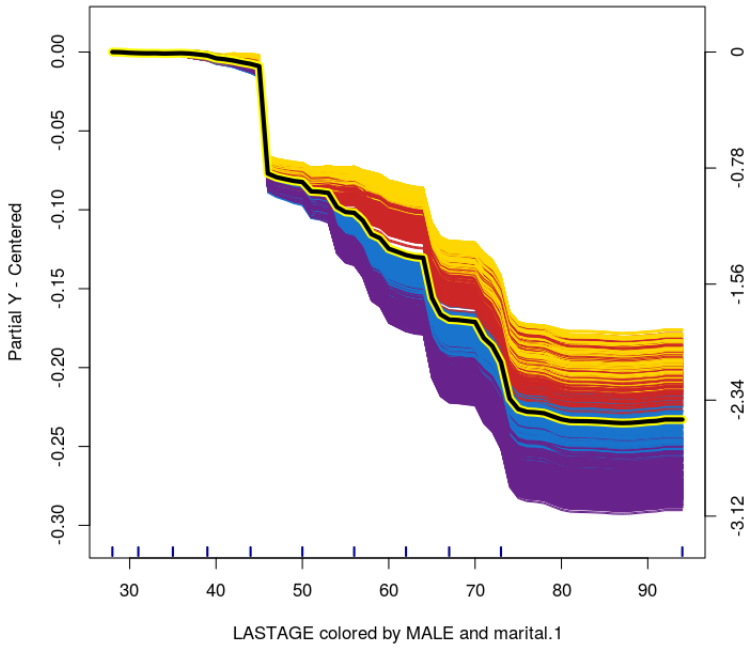

Figure 4.56: *Real Data Analysis - Centered ICE Plot for LASTAGE. If marital.1=0 and MALE=0, the ICE curve is red. If marital.1=0 and MALE=1, the ICE curve is blue. If marital.1=1 and MALE=0, the ICE curve is yellow. If marital.1=1 and MALE=1, the ICE curve is purple.*

Chapter 5

Conclusions

5.1 Discussion

We introduced a full one-step procedure to identify treatment effect heterogeneity by using ICE plots. We have corroborated Hahn *et al.* (2018b) study, which argues that inclusion of the propensity score can suppress at least part of the bias that the RIC phenomenon adds to the data. This idea was enforced by analyzing the effects of the propensity score through the use of ICE plots as a kind of sensitivity analysis, and by the use of a full-Bayesian variable selection method. The intuition behind the use of the propensity score as a covariate on tree-based models is that since elements with the same value of the propensity score are comparable, then the propensity score is actually partitioning the data between comparable individuals. As shown on Chapter 3, the treatment effects estimated among these partitions can be used to estimate the CATE. Since the propensity score already partitioned the data, the BART model can use this covariate to create splitting rules, while using the other covariates in order to capture traits for every group of individuals that belongs to the sample, allowing ITE estimates with less uncertainty than what was originally proposed by Hill (2011).

We have also found that in binary treatment effect observational studies even a naively estimated propensity score (which was played by the GLM in the simulations) may have a positive impact on the model, and if the estimates are completely random (like in the Rand models in the simulations), there might be an additional bias in the treatment effect estimation. However, these results were based on simulations, so *strong ignorability* was assured. As discussed in Section 3.4, this assumption cannot be guaranteed even in simple studies, so the conclusions based on the use of the propensity score must be handled carefully. Alternative Bayesian or machine learning estimations of the propensity score can also be further explored.

In regard of model effectiveness, the BCF allows its priors to be allocated freely in the functions

related to the prognostic effect and the treatment effect, so the model may held better results than BART if cross-validation is applied, but we have not found a clear superior model under default priors, hyperparameters and hyperpriors.

In general the ICE plots seemed to adequately capture the most important traits on models which the propensity score was correctly specified. This tool has proved to be incredibly useful in the causal inference setting and since it might be adapted to any model with ease, we advocate its use. Also, the c-ICE and d-ICE plots are able to indicate where probable interactions occur among the covariates.

The full-Bayesian variable selection also has proved to be useful since bias reduction was observed in the simulation studies. This feature might be useful in many situations on the causal inference setting, especially when $p > n$ and there is great uncertainty associated with the data.

It is important to note that one issue that was found in this study was the difficulty that some models had to achieve convergence. This may be due to the fact that only two Metropolis-Hastings proposals (GROW and PRUNE) are implemented on the *BART package*. This issue have affected especially the probit-BART model. The inclusion of other proposals, like the ones introduced by Pratola (2016), might be able to solve this issue.

Furthermore, it should be addressed that *strong ignorability* cannot be assured on any observational study, so any decisions that end up being made by using the Potential Outcomes framework must be treated very carefully, since unmeasured confounders, along with the inclusion of irrelevant covariates, may end up increasing the bias.

5.2 Possible Extensions

A possible extension of this study can be done by applying the DART Dirichlet hyperprior to the BCF model and verifying the model effectiveness in high dimensional data examples with $p > n$. Another possible approach could be done by inserting heteroscedastic error terms and applying Pratola *et al.* (2017) approach.

Lastly, the treatment effect data generation in this study followed Hahn *et al.* (2018b) example, meaning that this study can be expanded to other data generation processes in order to assess the use of the tools proposed.

Appendix A

Supplementary Material

Let y_{ij} denote the jth observation in the terminal node $i \in \{1, 2, ..., b\}$ in a tree with structure T. It will be assumed that

$$y_{i1}, y_{i2}, ..., y_{in_i} \overset{iid}{\sim} \mathcal{N}\left(\mu_i, \sigma^2\right), \qquad \mu_1, \mu_2, ..., \mu_b \overset{iid}{\sim} \mathcal{N}\left(k, \sigma_\mu^2\right),$$

such that $n_1 + n_2 + ... + n_b = n$, and $M = \{\mu_1, ..., \mu_b\}$.

So, the likelihood of the data following Chipman *et al.* (1998) framework is given by

$$p\left(\tilde{y} \mid \mathbf{x}, T, M, \sigma^2\right) = \prod_{i=1}^{b} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_{ij} - \mu_i\right)^2}{2\sigma^2}\right).$$

In order to avoid reversible jumps, the MCMC algorithm uses

$$p\left(\tilde{y} \mid \mathbf{x}, T, \sigma\right) = \int p\left(\tilde{y} \mid \mathbf{x}, T, M, \sigma^2\right) p\left(M \mid T, \sigma\right) dM, \qquad p\left(M \mid T, \sigma\right) = \prod_{i=1}^{b} p\left(\mu_i \mid T, \sigma\right).$$

So the integral can be rewritten as

$$p\left(\tilde{y} \mid \mathbf{x}, T, \sigma\right) = \int ... \int p\left(\tilde{y} \mid \mathbf{x}, T, M, \sigma^2\right) \prod_{i=1}^{b} p\left(\mu_i \mid T, \sigma\right) d\mu_1 ... d\mu_b, \tag{A.1}$$

which means that the integral for each μ_i only have non-constant terms at the ith term of the product over the i index. In order to simplify, let us analyze only the integral over μ_1.

$$\int \prod_{i=1}^{b} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_{ij} - \mu_i\right)^2}{2\sigma^2}\right) \prod_{i=1}^{b} p\left(\mu_i \mid T, \sigma\right) d\mu_1 =$$

$$\int \prod_{i=1}^{b} \left(\sqrt{2\pi\sigma^2}\right)^{-\frac{n_i}{2}} \exp\left(-\frac{\sum_{j=1}^{n_i}\left(y_{ij} - \mu_i\right)^2}{2\sigma^2}\right) \prod_{i=1}^{b} p\left(\mu_i \mid T, \sigma\right) d\mu_1 =$$

$$\int \prod_{i=1}^{b} \left(\sqrt{2\pi\sigma^2}\right)^{-\frac{n_i}{2}} \exp\left(-\frac{\sum_{j=1}^{n_i}\left(y_{ij} - \mu_i\right)^2}{2\sigma^2}\right) \left(\sqrt{2\pi\sigma_\mu^2}\right)^{-\frac{1}{2}} \exp\left(-\frac{\left(\mu_i - k\right)^2}{2\sigma_\mu^2}\right) d\mu_1 =$$

$$\int \prod_{i=1}^{b} \left(\sqrt{2\pi\sigma^2}\right)^{-\frac{n_i}{2}} \left(\sqrt{2\pi\sigma_\mu^2}\right)^{-\frac{1}{2}} \exp\left(-\frac{\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2}{2\sigma^2} - \frac{(\mu_i-k)^2}{2\sigma_\mu^2}\right) d\mu_1 =$$

$$\begin{aligned} &\prod_{i=2}^{b} \left(\sqrt{2\pi\sigma^2}\right)^{-\frac{n_i}{2}} \left(\sqrt{2\pi\sigma_\mu^2}\right)^{-\frac{1}{2}} \exp\left(-\frac{\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2}{2\sigma^2} - \frac{(\mu_i-k)^2}{2\sigma_\mu^2}\right) \\ &\int \left(\sqrt{2\pi\sigma^2}\right)^{-\frac{n_1}{2}} \left(\sqrt{2\pi\sigma_\mu^2}\right)^{-\frac{1}{2}} \exp\left(-\frac{\sum_{j=1}^{n_1}(y_{1j}-\mu_1)^2}{2\sigma^2} - \frac{(\mu_1-k)^2}{2\sigma_\mu^2}\right) d\mu_1. \end{aligned} \tag{A.2}$$

Solving only the integral, we have

$$\int \left(\sqrt{2\pi\sigma^2}\right)^{-\frac{n_1}{2}} \left(\sqrt{2\pi\sigma_\mu^2}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n_1}(y_{1j}-\mu_1)^2 - \frac{1}{2\sigma_\mu^2}(\mu_1-k)^2\right) d\mu_1 =$$

$$\begin{aligned} &\left(\sqrt{2\pi\sigma^2}\right)^{-\frac{n_1}{2}} \left(\sqrt{2\pi\sigma_\mu^2}\right)^{-\frac{1}{2}} \\ &\int \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{n_1} y_{1j}^2 - 2\mu_1\sum_{j=1}^{n_1} y_{1j} + n_1\mu_1^2\right) - \frac{1}{2\sigma_\mu^2}\left(\mu_1^2 - 2\mu_1 k + k^2\right)\right) d\mu_1 = \end{aligned}$$

$$\begin{aligned} &\left(\sqrt{2\pi\sigma^2}\right)^{-\frac{n_1}{2}} \left(\sqrt{2\pi\sigma_\mu^2}\right)^{-\frac{1}{2}} \exp\left(-\frac{\sum_{j=1}^{n_1} y_{1j}^2}{2\sigma^2} - \frac{k^2}{2\sigma_\mu^2}\right) \\ &\int \exp\left(-\frac{1}{2\sigma^2}\left(-2\mu_1\sum_{j=1}^{n_1} y_{1j} + n_1\mu_1^2\right) - \frac{1}{2\sigma_\mu^2}\left(\mu_1^2 - 2\mu_1 k\right)\right) d\mu_1. \end{aligned} \tag{A.3}$$

Again, solving only the integral, we have

$$\int \exp\left(-\frac{1}{2\sigma^2}\left(-2\mu_1\sum_{j=1}^{n_1} y_{1j} + n_1\mu_1^2\right) - \frac{1}{2\sigma_\mu^2}\left(\mu_1^2 - 2\mu_1 k\right)\right) d\mu_1 =$$

$$\int \exp\left(-\frac{1}{4\sigma^2\sigma_\mu^2}\left(-4\sigma_\mu^2\mu_1\sum_{j=1}^{n_1} y_{1j} + 2\sigma_\mu^2 n_1\mu_1^2 + 2\sigma^2\mu_1^2 - 4\sigma^2\mu_1 k\right)\right) d\mu_1 =$$

$$\int \exp\left(-\frac{1}{4\sigma^2\sigma_\mu^2}\left(\left(-4\sigma_\mu^2\sum_{j=1}^{n_1} y_{1j} - 4\sigma^2 k\right)\mu_1 + \left(2\sigma_\mu^2 n_1 + 2\sigma^2\right)\mu_1^2\right)\right) d\mu_1 =$$

$$\int \exp\left(-\frac{1}{\frac{4\sigma^2\sigma_\mu^2}{2\left(\sigma_\mu^2 n_1+\sigma^2\right)}}\left(-2\frac{\left(2\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}+2\sigma^2 k\right)}{2\left(\sigma_\mu^2 n_1+\sigma^2\right)}\mu_1+\mu_1^2\right)\right)d\mu_1 =$$

$$\int \exp\left(-\frac{1}{2\left(\frac{\sigma^2\sigma_\mu^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)}\left(-2\left(\frac{\left(2\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}+2\sigma^2 k\right)}{2\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)\mu_1+\mu_1^2\right)\right)d\mu_1.$$

It is possible to notice that this integral looks like a normal distribution.

$$\left(2\pi\left(\frac{\sigma^2\sigma_\mu^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)\right)^{\frac{1}{2}}\exp\left(\frac{\left(\frac{\left(\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}+\sigma^2 k\right)}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^2}{2\left(\frac{\sigma^2\sigma_\mu^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)}\right)\int\left(2\pi\left(\frac{\sigma^2\sigma_\mu^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)\right)^{-\frac{1}{2}}$$
$$\exp\left(-\frac{\left(\left(\frac{\left(\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}+\sigma^2 k\right)}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^2-2\left(\frac{\left(\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}+\sigma^2 k\right)}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)\mu_1+\mu_1^2\right)}{2\left(\frac{2\sigma^2\sigma_\mu^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)}\right)d\mu_1 =$$

$$\left(2\pi\left(\frac{\sigma^2\sigma_\mu^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)\right)^{\frac{1}{2}}\exp\left(\frac{\left(\frac{\left(\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}+\sigma^2 k\right)}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^2}{2\left(\frac{\sigma^2\sigma_\mu^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)}\right) \tag{A.4}$$

By replacing (A.4) in (A.3), we have,

$$\left(\sqrt{2\pi\sigma^2}\right)^{-\frac{n_1}{2}}\left(\sqrt{2\pi\sigma_\mu^2}\right)^{-\frac{1}{2}}\exp\left(-\frac{\sum_{j=1}^{n_1}y_{1j}^2}{2\sigma^2}-\frac{k^2}{2\sigma_\mu^2}\right)\left(2\pi\left(\frac{\sigma^2\sigma_\mu^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)\right)^{\frac{1}{2}}$$
$$\exp\left(\frac{\left(\frac{\left(\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}+\sigma^2 k\right)}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^2}{2\left(\frac{\sigma^2\sigma_\mu^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(2\pi\sigma_\mu^2\right)^{-\frac{1}{2}}\left(2\pi\sigma_\mu^2\right)^{\frac{1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}$$
$$\exp\left(-\frac{\sum_{j=1}^{n_1}y_{1j}^2}{2\sigma^2}-\frac{k^2}{2\sigma_\mu^2}+\frac{\left(\frac{\left(\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}+\sigma^2 k\right)}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^2}{2\left(\frac{\sigma^2\sigma_\mu^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}\exp\left(-\frac{\sum_{j=1}^{n_1}y_{1j}^2}{2\sigma^2}-\frac{k^2}{2\sigma_\mu^2}+\frac{\left(\frac{\left(\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}+\sigma^2k\right)}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^2}{2\left(\frac{\sigma^2\sigma_\mu^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}\exp\left(-\frac{\sum_{j=1}^{n_1}y_{1j}^2}{2\sigma^2}-\frac{k^2}{2\sigma_\mu^2}+\frac{\frac{\left(\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}+\sigma^2k\right)^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}}{2\left(\sigma^2\sigma_\mu^2\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}$$
$$\exp\left(-\frac{\sum_{j=1}^{n_1}y_{1j}^2}{2\sigma^2}-\frac{k^2}{2\sigma_\mu^2}+\frac{\left(\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}\right)^2+2\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}\sigma^2k+\left(\sigma^2k\right)^2}{2\sigma^2\sigma_\mu^2\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}$$
$$\exp\left(-\frac{\sum_{j=1}^{n_1}y_{1j}^2}{2\sigma^2}-\frac{k^2\sigma^2\left(\sigma_\mu^2 n_1+\sigma^2\right)}{2\sigma^2\sigma_\mu^2\left(\sigma_\mu^2 n_1+\sigma^2\right)}+\frac{\left(\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}\right)^2+2\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}\sigma^2k+\left(\sigma^2k\right)^2}{2\sigma^2\sigma_\mu^2\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}$$
$$\exp\left(-\frac{\sum_{j=1}^{n_1}y_{1j}^2}{2\sigma^2}+\frac{\left(\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}\right)^2+2\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}\sigma^2k+\left(\sigma^2k\right)^2-k^2\sigma^2\sigma_\mu^2n_1-\left(k\sigma^2\right)^2}{2\sigma^2\sigma_\mu^2\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}$$
$$\exp\left(-\frac{\sum_{j=1}^{n_1}y_{1j}^2}{2\sigma^2}+\frac{\left(\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}\right)^2+2\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}\sigma^2k-k^2\sigma^2\sigma_\mu^2n_1}{2\sigma^2\sigma_\mu^2\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}\exp\left(-\frac{\sum_{j=1}^{n_1}y_{1j}^2}{2\sigma^2}\right)+$$

$$\exp\left(\frac{\sigma_\mu^2\left(\sigma_\mu^2 n_1+\sigma^2\right)n_1\left(\frac{\sum_{j=1}^{n_1}y_{1j}}{n_1}\right)^2-\sigma_\mu^2\sigma^2 n_1\left(\frac{\sum_{j=1}^{n_1}y_{1j}}{n_1}\right)^2+2\sigma_\mu^2\sum_{j=1}^{n_1}y_{1j}\sigma^2 k-k^2\sigma^2\sigma_\mu^2 n_1}{2\sigma^2\sigma_\mu^2\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}$$

$$\exp\left(-\frac{\sum_{j=1}^{n_1}y_{1j}^2}{2\sigma^2}+\frac{n_1\left(\frac{\sum_{j=1}^{n_1}y_{1j}}{n_1}\right)^2}{2\sigma^2}-\frac{n_1\left(\frac{\sum_{j=1}^{n_1}y_{1j}}{n_1}\right)^2-2n_1\left(\frac{\sum_{j=1}^{n_1}y_{1j}}{n_1}\right)k+k^2n_1}{2\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}$$

$$\exp\left(-\frac{\sum_{j=1}^{n_1}y_{1j}^2}{2\sigma^2}+\frac{n_1\left(\frac{\sum_{j=1}^{n_1}y_{1j}}{n_1}\right)^2}{2\sigma^2}-n_1\frac{\left(\frac{\sum_{j=1}^{n_1}y_{1j}}{n_1}\right)^2-2\left(\frac{\sum_{j=1}^{n_1}y_{1j}}{n_1}\right)k+k^2}{2\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}$$

$$\exp\left(-\frac{\sum_{j=1}^{n_1}y_{1j}^2}{2\sigma^2}+\frac{n_1\left(\frac{\sum_{j=1}^{n_1}y_{1j}}{n_1}\right)^2}{2\sigma^2}-\frac{n_1\left(\left(\frac{\sum_{j=1}^{n_1}y_{1j}}{n_1}\right)-k\right)^2}{2\left(n_1\sigma_\mu^2+\sigma^2\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}$$

$$\exp\left(-\left(\frac{\sum_{j=1}^{n_1}y_{1j}^2-n_1\left(\frac{\sum_{j=1}^{n_1}y_{1j}}{n_1}\right)^2}{2\sigma^2}\right)-\frac{n_1\left(\left(\frac{\sum_{j=1}^{n_1}y_{1j}}{n_1}\right)-k\right)^2}{2\left(n_1\sigma_\mu^2+\sigma^2\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}$$
$$\exp\left(-\left(\frac{\sum_{j=1}^{n_1} y_{1j}^2-2n_1\left(\frac{\sum_{j=1}^{n_1} y_{1j}}{n_1}\right)^2+n_1\left(\frac{\sum_{j=1}^{n_1} y_{1j}}{n_1}\right)^2}{2\sigma^2}\right)-\frac{n_1\left(\left(\frac{\sum_{j=1}^{n_1} y_{1j}}{n_1}\right)-k\right)^2}{2\left(n_1\sigma_\mu^2+\sigma^2\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}$$
$$\exp\left(-\left(\frac{\sum_{j=1}^{n_1} y_{1j}^2-2\sum_{j=1}^{n_1} y_{1j}\left(\frac{\sum_{j=1}^{n_1} y_{1j}}{n_1}\right)+n_1\left(\frac{\sum_{j=1}^{n_1} y_{1j}}{n_1}\right)^2}{2\sigma^2}\right)-\frac{n_1\left(\left(\frac{\sum_{j=1}^{n_1} y_{1j}}{n_1}\right)-k\right)^2}{2\left(n_1\sigma_\mu^2+\sigma^2\right)}\right)=$$

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}\exp\left(-\frac{\sum_{j=1}^{n_1}\left(y_{1j}+\left(\frac{\sum_{j=1}^{n_1} y_{1j}}{n_1}\right)\right)^2}{2\sigma^2}-\frac{n_1\left(\left(\frac{\sum_{j=1}^{n_1} y_{1j}}{n_1}\right)-k\right)^2}{2\left(n_1\sigma_\mu^2+\sigma^2\right)}\right).$$

Now we can substitute this equation in (A.2) to obtain

$$\left(2\pi\sigma^2\right)^{-\frac{n_1}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_1+\sigma^2\right)}\right)^{\frac{1}{2}}\exp\left(-\frac{\sum_{j=1}^{n_1}\left(y_{1j}+\left(\frac{\sum_{j=1}^{n_1} y_{1j}}{n_1}\right)\right)^2}{2\sigma^2}-\frac{n_1\left(\left(\frac{\sum_{j=1}^{n_1} y_{1j}}{n_1}\right)-k\right)^2}{2\left(n_1\sigma_\mu^2+\sigma^2\right)}\right)$$
$$\prod_{i=2}^{b}\left(\sqrt{2\pi\sigma^2}\right)^{-\frac{n_i}{2}}\left(\sqrt{2\pi\sigma_\mu^2}\right)^{-\frac{1}{2}}\exp\left(-\frac{\sum_{j=1}^{n_i}\left(y_{ij}-\mu_i\right)^2}{2\sigma^2}-\frac{\left(\mu_i-k\right)^2}{2\sigma_\mu^2}\right).$$

By using the same argument to integrate the variables $\mu_2, ..., \mu_b$ in (A.1), we can show that

$$p\left(\tilde{y}\mid\mathbf{x},T,\sigma\right)=$$
$$\prod_{i=1}^{b}\left(2\pi\sigma^2\right)^{-\frac{n_i}{2}}\left(\frac{\sigma^2}{\left(\sigma_\mu^2 n_i+\sigma^2\right)}\right)^{\frac{1}{2}}\exp\left(-\frac{\sum_{j=1}^{n_i}\left(y_{ij}+\left(\frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}\right)\right)^2}{2\sigma^2}-\frac{n_i\left(\left(\frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}\right)-k\right)^2}{2\left(n_i\sigma_\mu^2+\sigma^2\right)}\right).$$

This is the same result presented by Linero (2017), since the author shows that the likelihood can be simplified to

$$p\left(\tilde{y}\mid\mathbf{x},T,\sigma\right)=\prod_{i=1}^{b}\left(2\pi\sigma^2\right)^{-\frac{n_i}{2}}\sqrt{\frac{\sigma^2}{n_i\sigma_\mu^2+\sigma^2}}\exp\left(-\frac{\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2}{2\sigma^2}-\frac{n_i\left(\bar{y}_i-k\right)^2}{2\left(n_i\sigma_\mu^2+\sigma^2\right)}\right),$$

where $\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$.

Appendix B

Graphics - Toy Example

B.1 Convergence Analysis

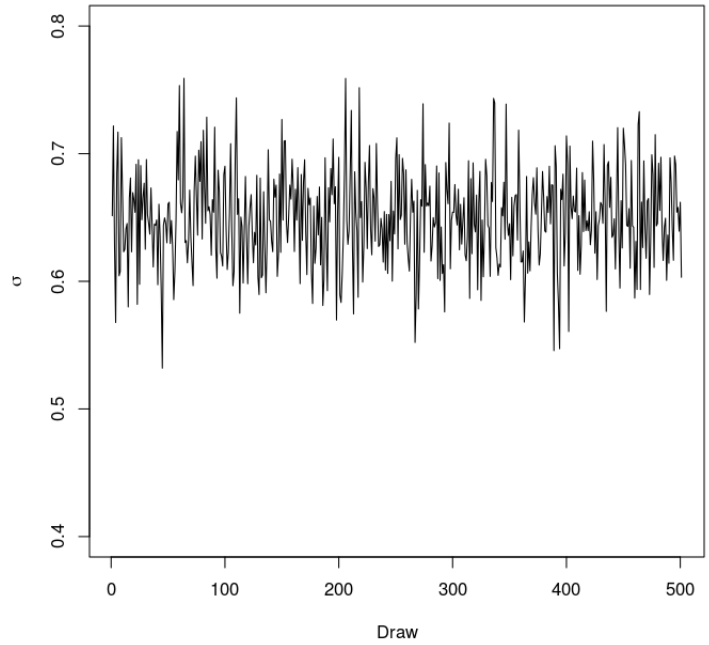

Figure B.1: *Vanilla model (Toy Example) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

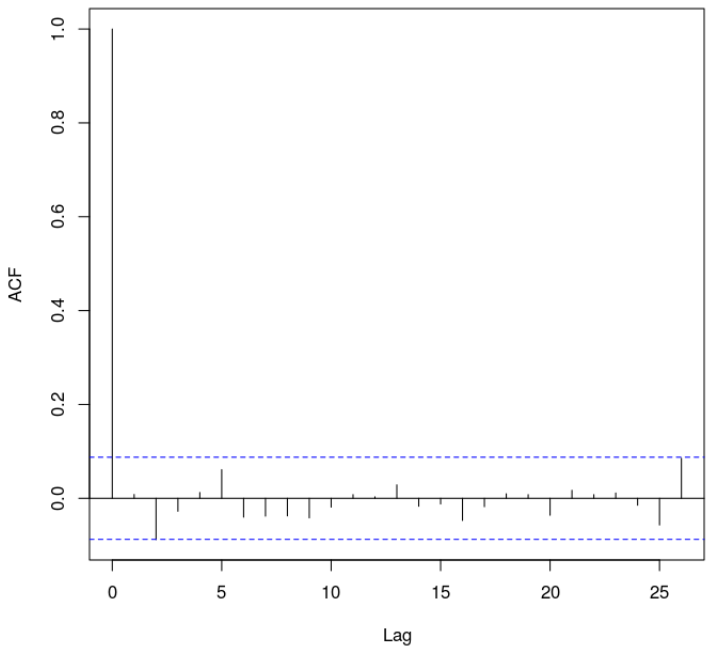

Figure B.2: *Vanilla model (Toy Example) - ACF function for the σ draws. Apparently the autocorrelation among the draws is low.*

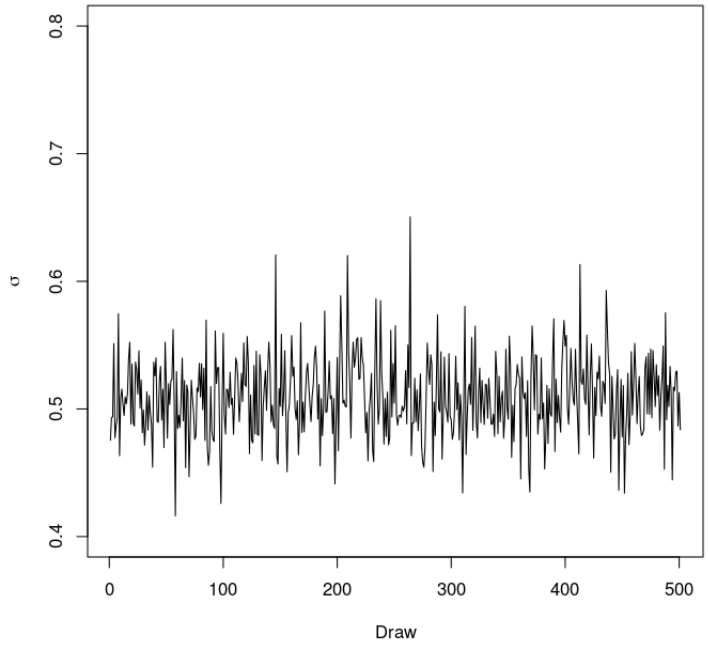

Figure B.3: *Oracle model (Toy Example) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

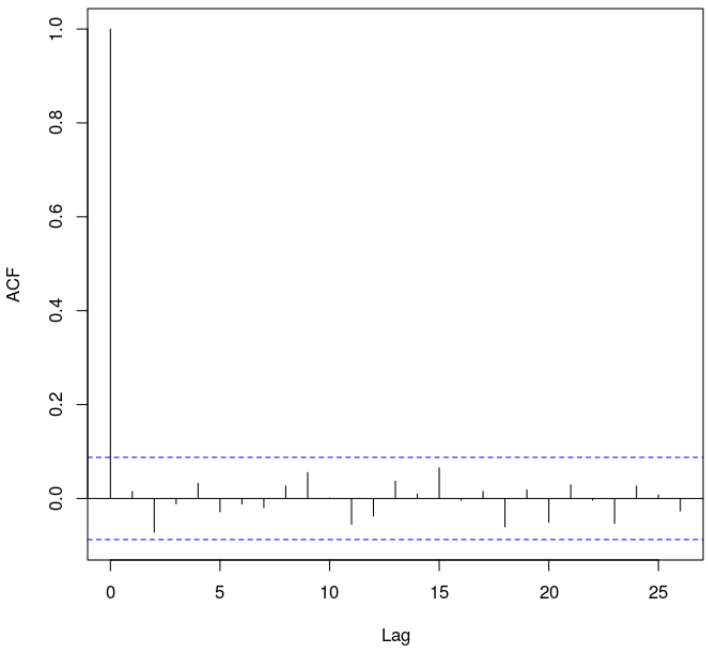

Figure B.4: *Oracle model (Toy Example) - ACF function for the σ draws. Apparently the autocorrelation among the draws is low.*

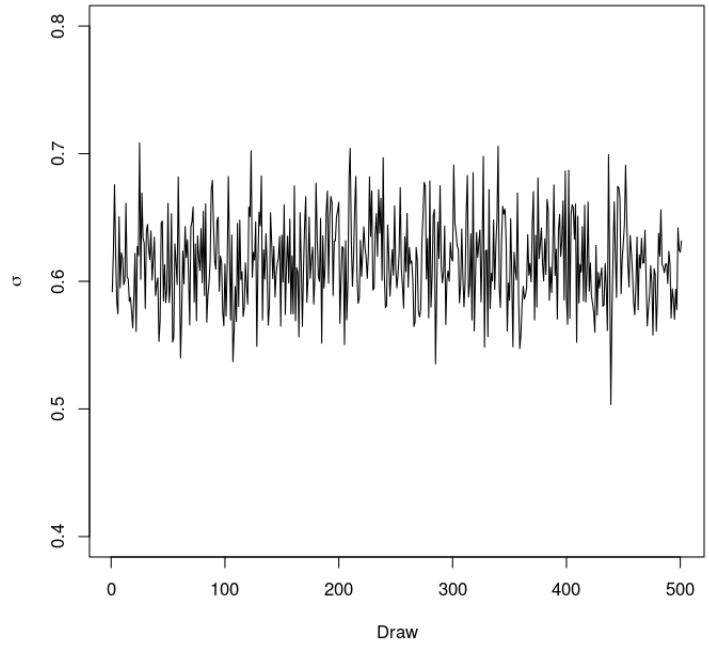

Figure B.5: *PS-BART model (Toy Example) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

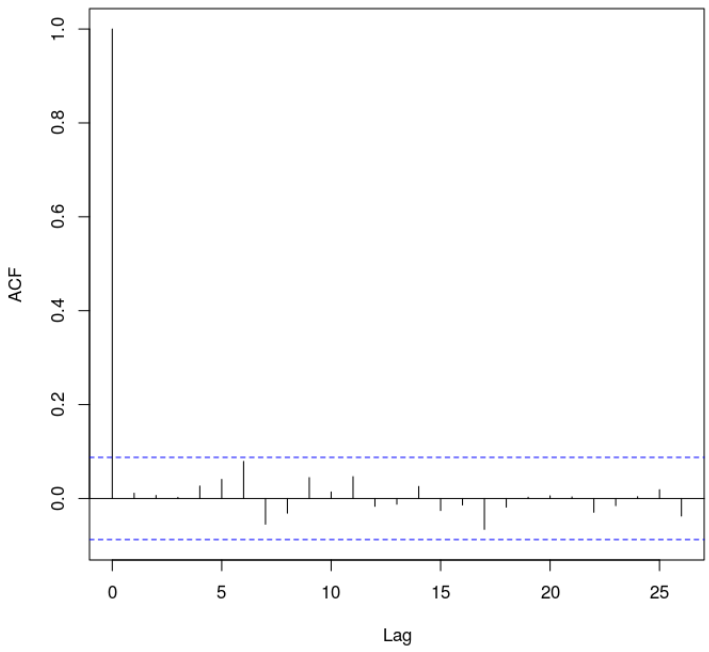

Figure B.6: *PS-BART model (Toy Example) - ACF function for the σ draws. Apparently the autocorrelation among the draws is low.*

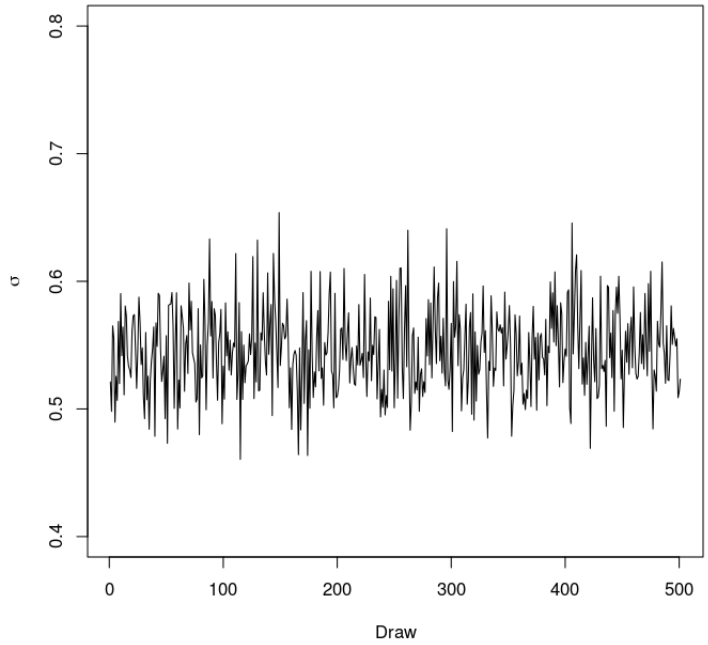

Figure B.7: *GLM-BART model (Toy Example) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

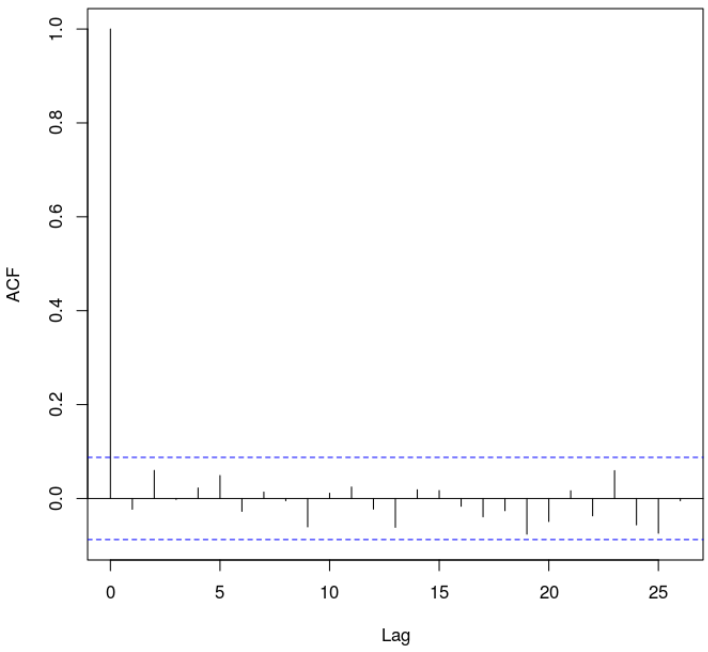

Figure B.8: *GLM-BART model (Toy Example) - ACF function for the σ draws. Apparently the autocorrelation among the draws is low.*

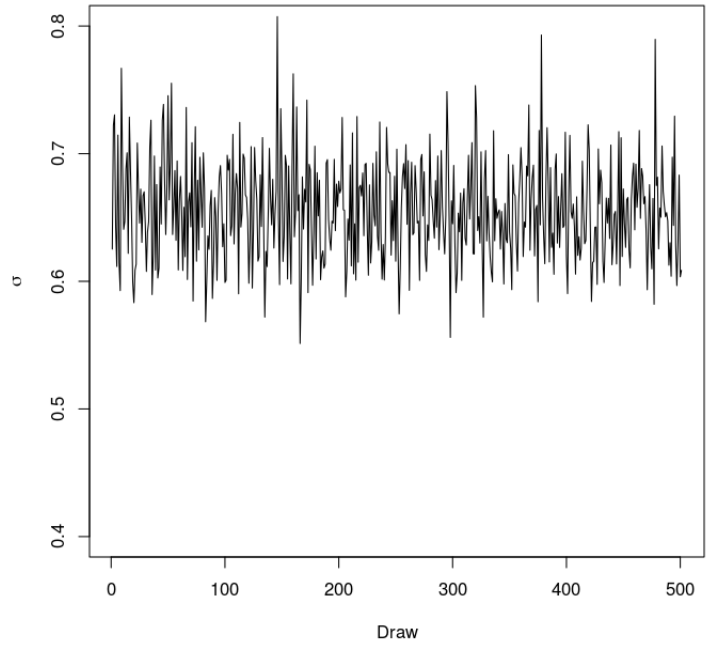

Figure B.9: *Rand-BART model (Toy Example) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

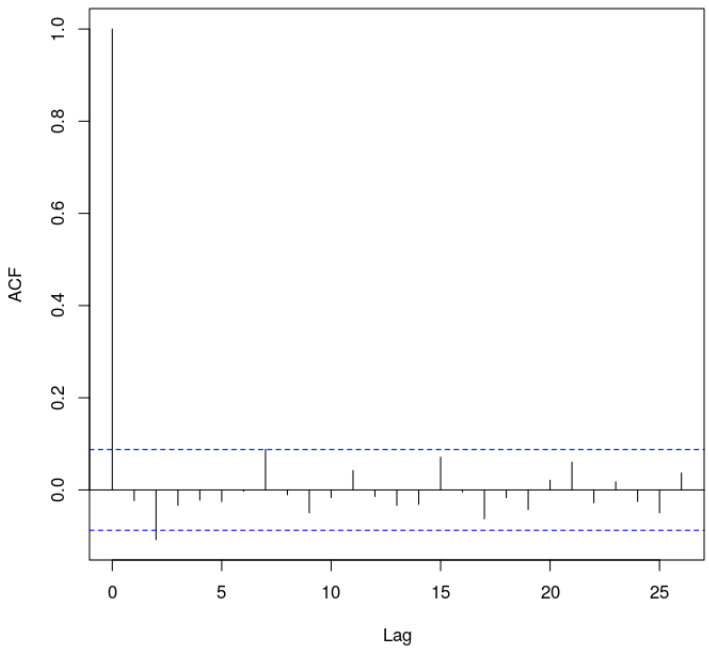

Figure B.10: *Rand-BART model (Toy Example) - ACF function for the σ draws. Apparently the autocorrelation among the draws is low.*

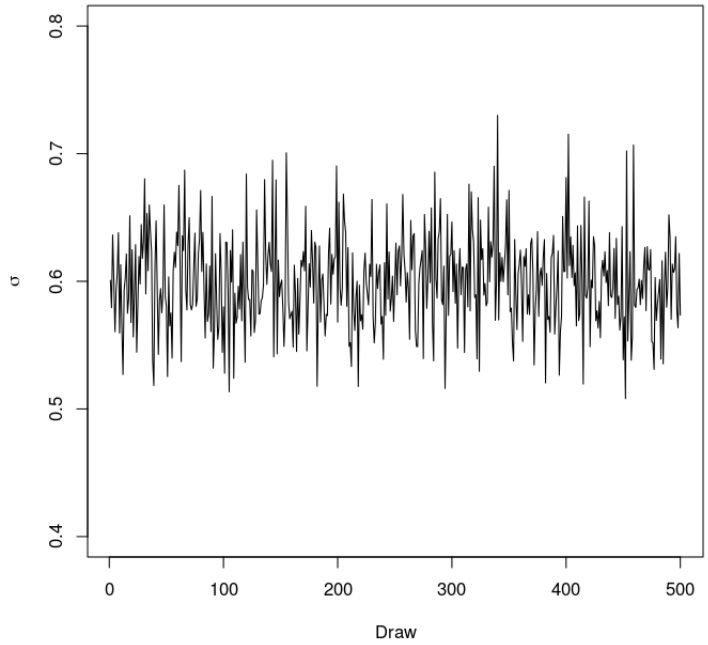

Figure B.11: *BART-BCF model (Toy Example) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

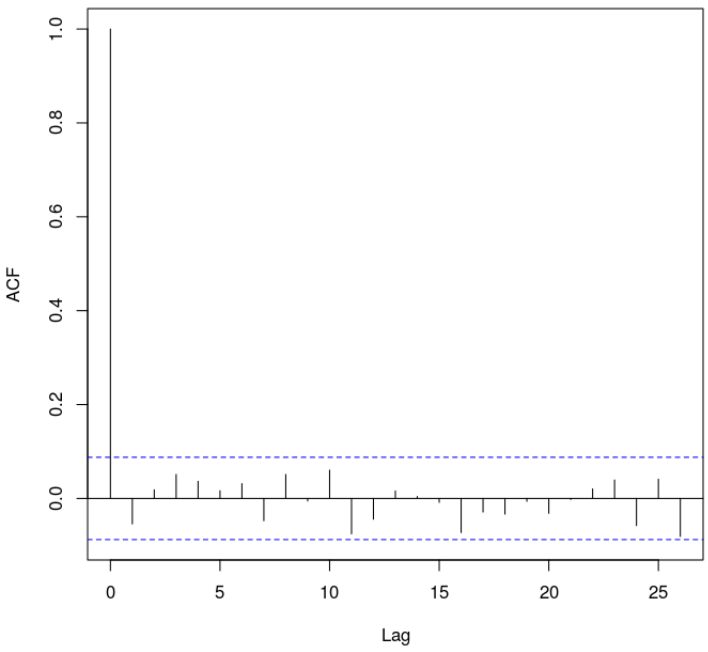

Figure B.12: *BART-BCF model (Toy Example) - ACF function for the σ draws. Apparently the autocorrelation among the draws is low.*

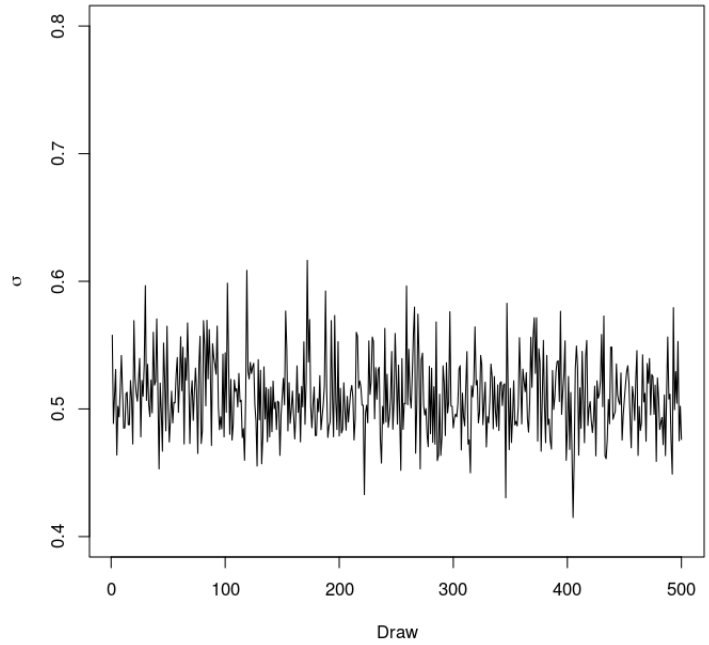

Figure B.13: *Oracle-BCF model (Toy Example) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

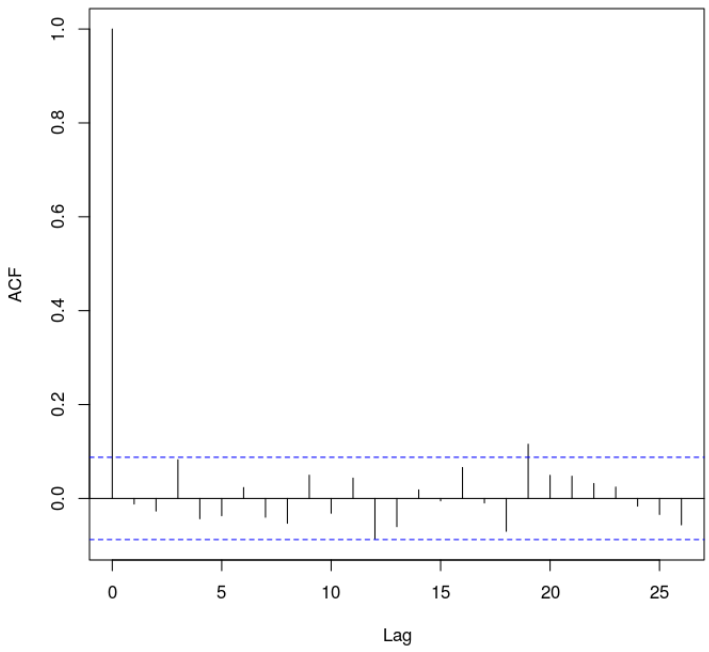

Figure B.14: *Oracle-BCF model (Toy Example) - ACF function for the σ draws. Apparently the autocorrelation among the draws is low.*

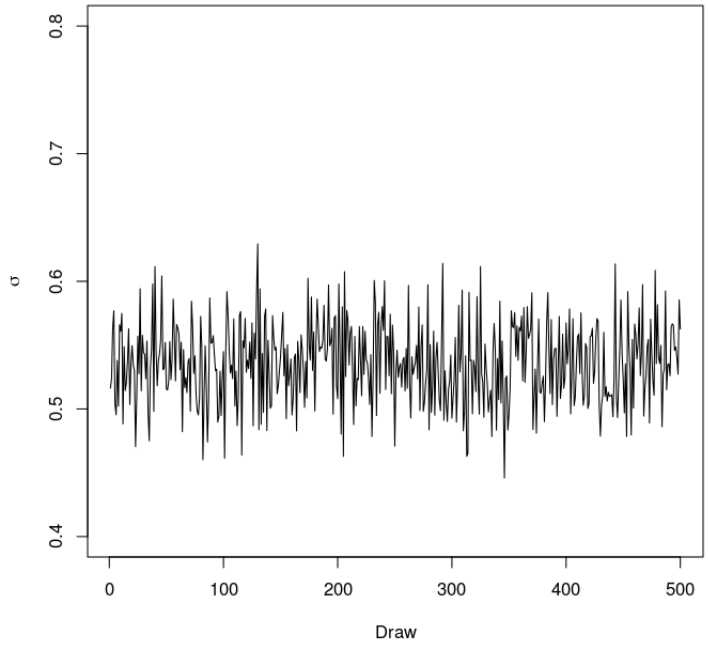

Figure B.15: *GLM-BCF model (Toy Example) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

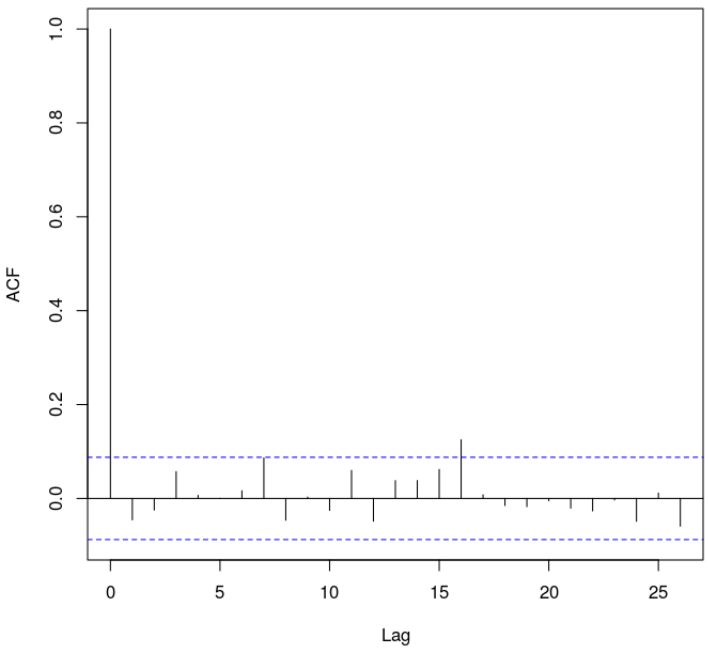

Figure B.16: *GLM-BCF model (Toy Example) - ACF function for the σ draws. Apparently the autocorrelation among the draws is low.*

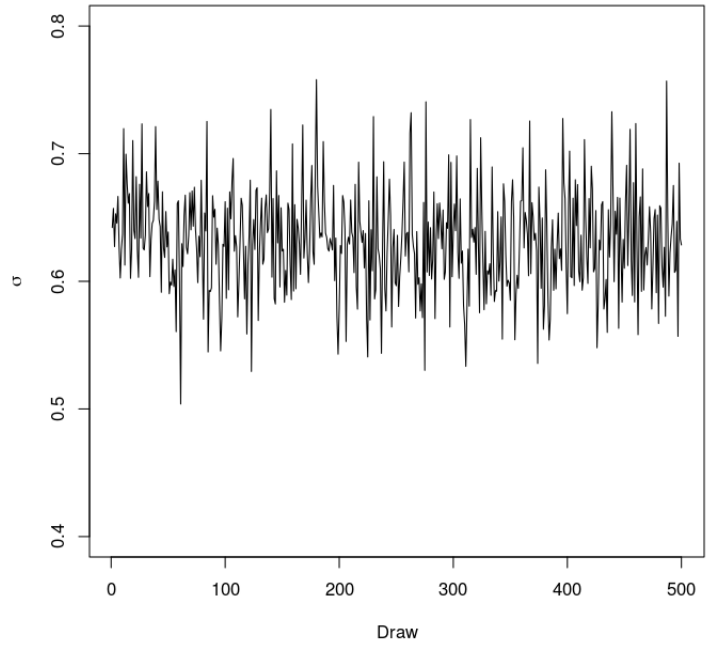

Figure B.17: *Rand-BCF model (Toy Example) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

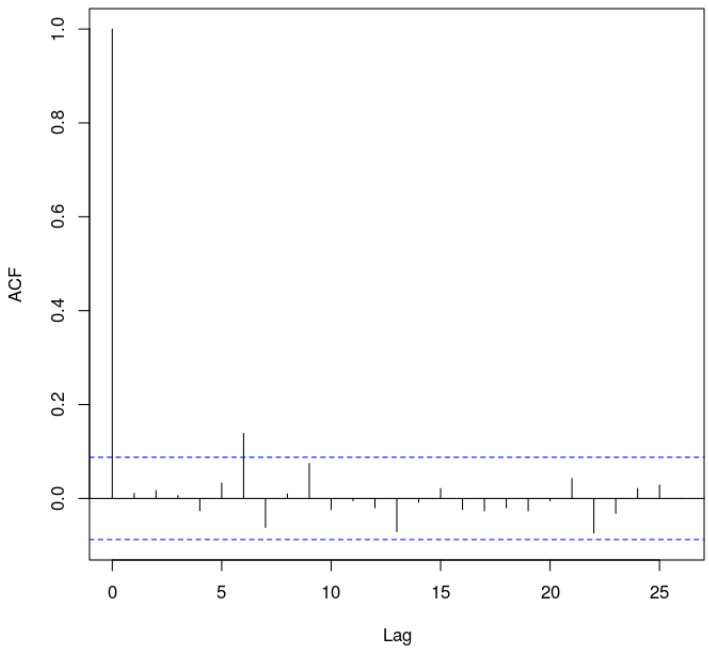

Figure B.18: *Rand-BCF model (Toy Example) - ACF function for the σ draws. Apparently the autocorrelation among the draws is low.*

B.2 ICE Plots

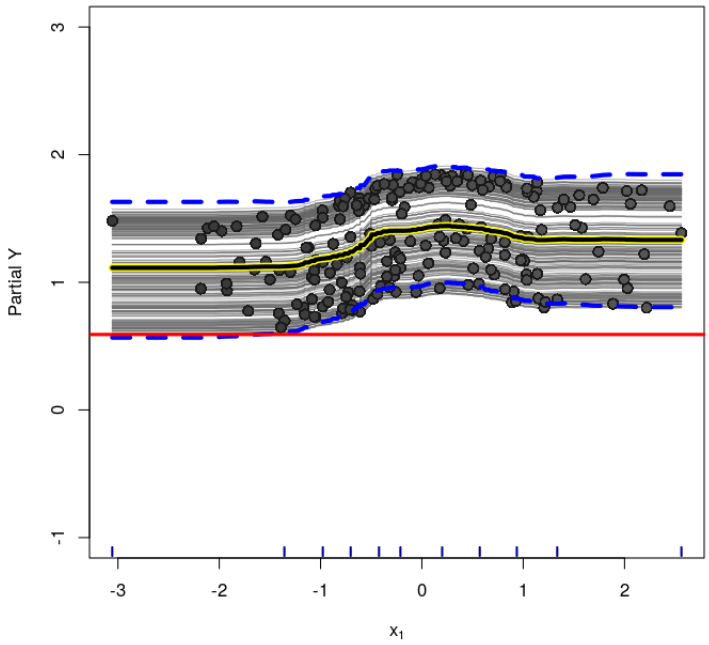

Figure B.19: *Vanilla model (Toy Example) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

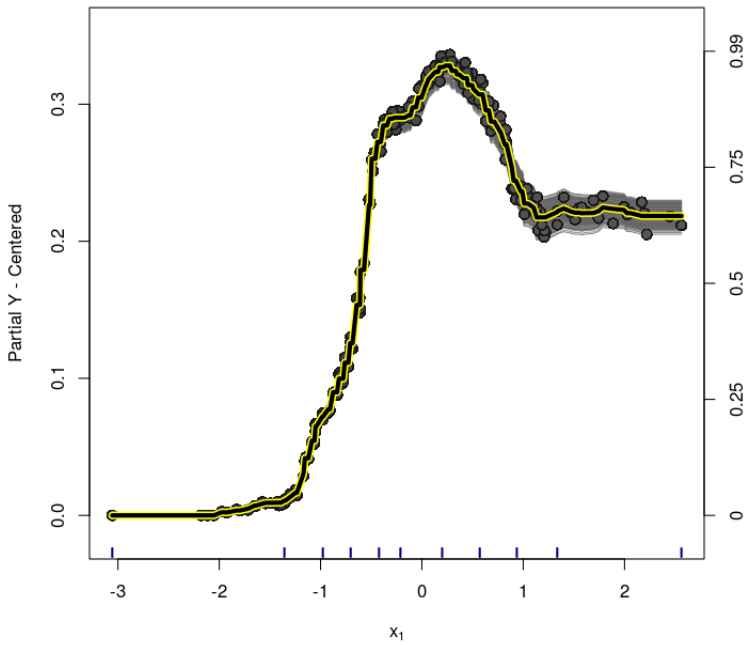

Figure B.20: *Vanilla model (Toy Example) - Variable x_1 - centered-ICE Plot for the treatment effect.*

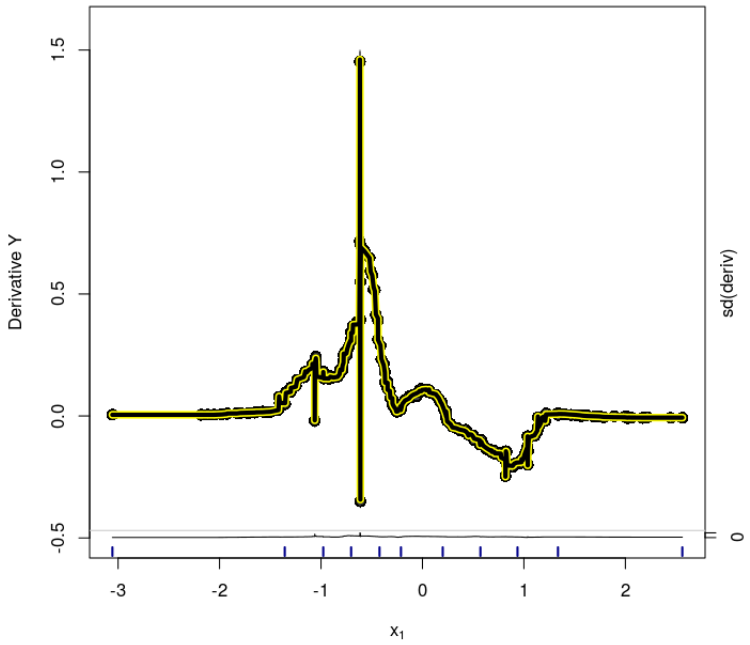

Figure B.21: *Vanilla model (Toy Example) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

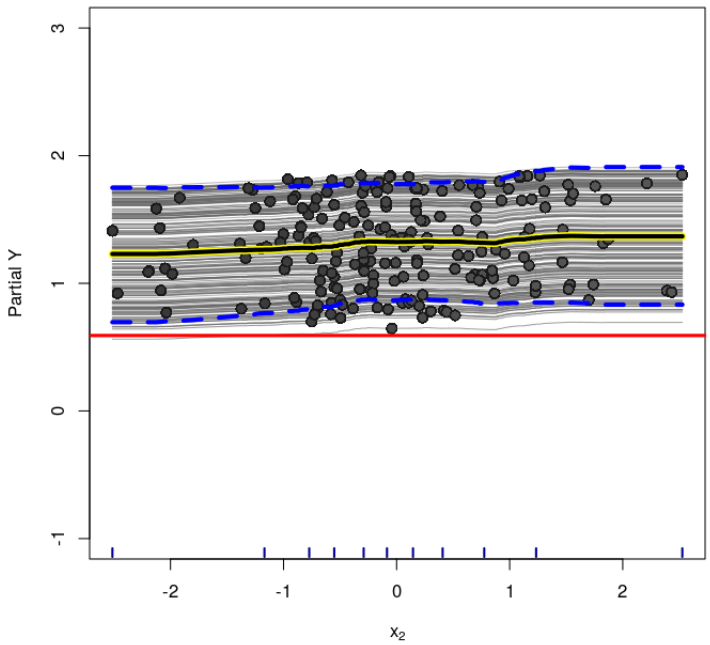

Figure B.22: *Vanilla model (Toy Example) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

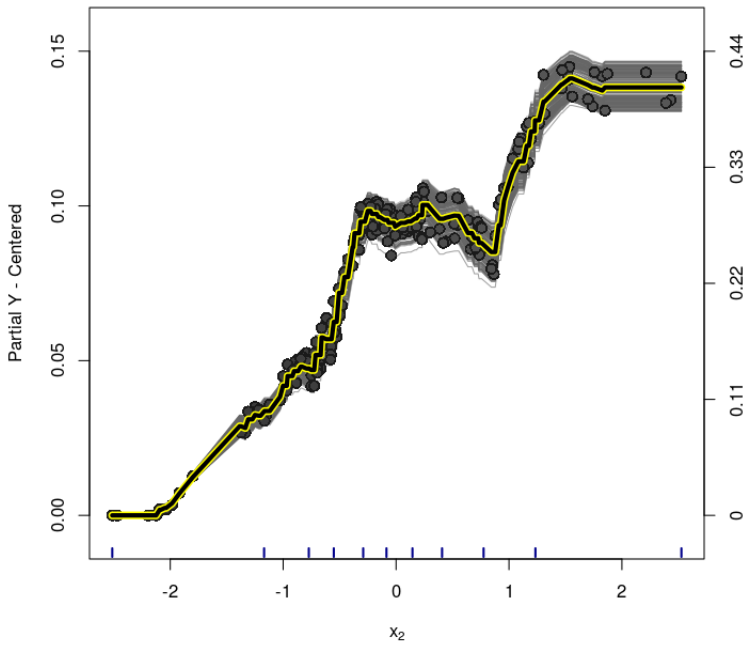

Figure B.23: *Vanilla model (Toy Example) - Variable x_2 - centered-ICE Plot for the treatment effect.*

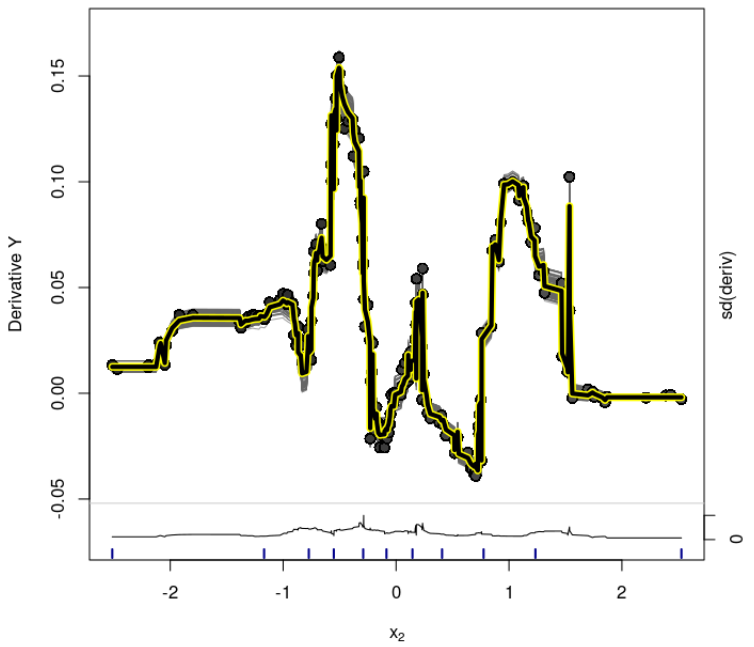

Figure B.24: *Vanilla model (Toy Example) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

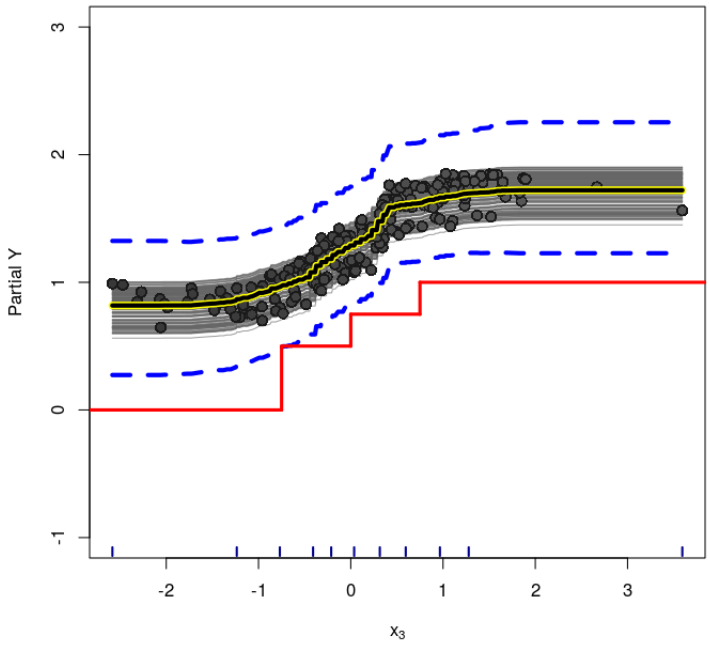

Figure B.25: *Vanilla model (Toy Example) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

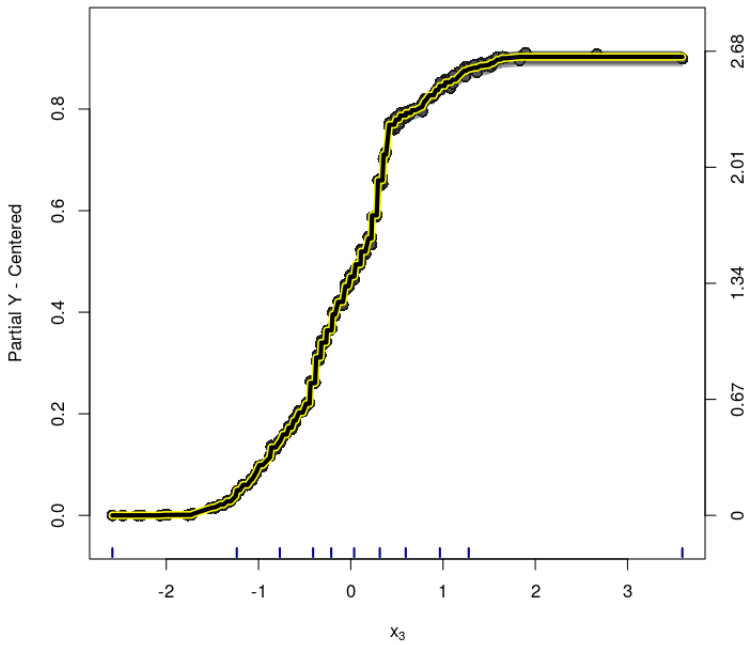

Figure B.26: *Vanilla model (Toy Example) - Variable x_3 - centered-ICE Plot for the treatment effect.*

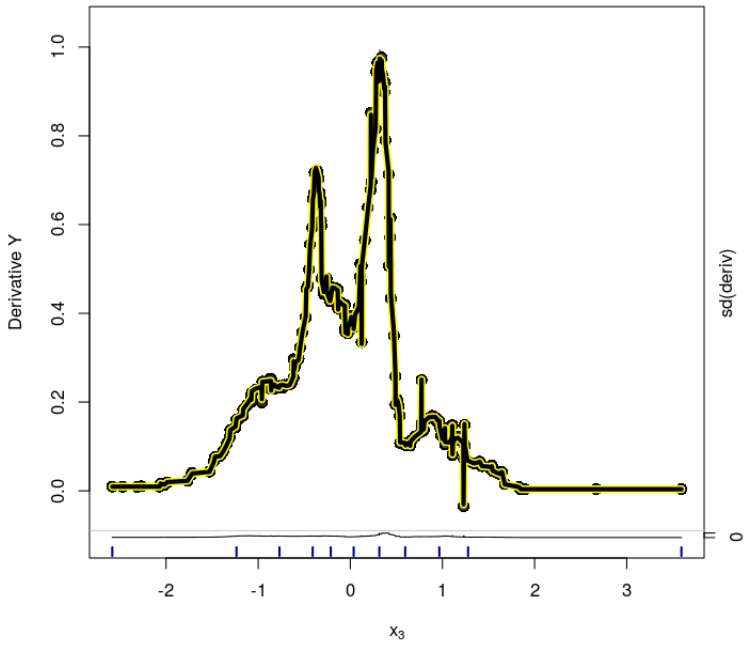

Figure B.27: *Vanilla model (Toy Example) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

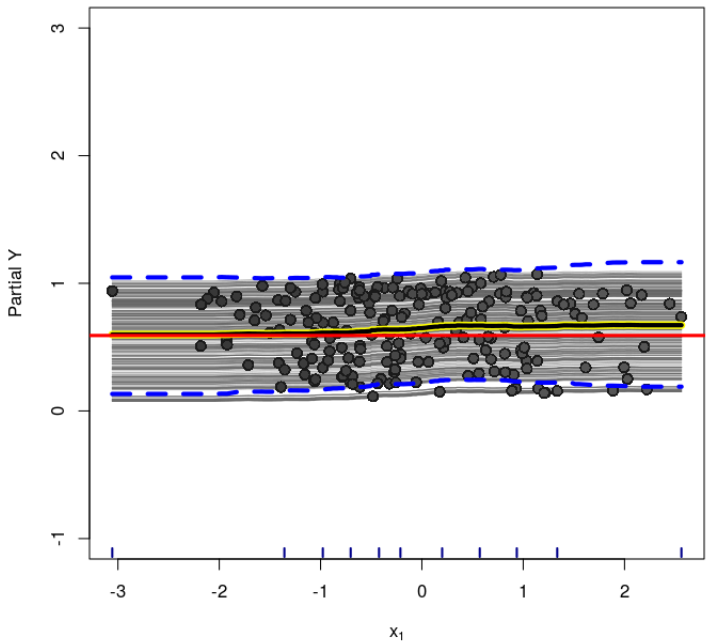

Figure B.28: *Oracle model (Toy Example) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

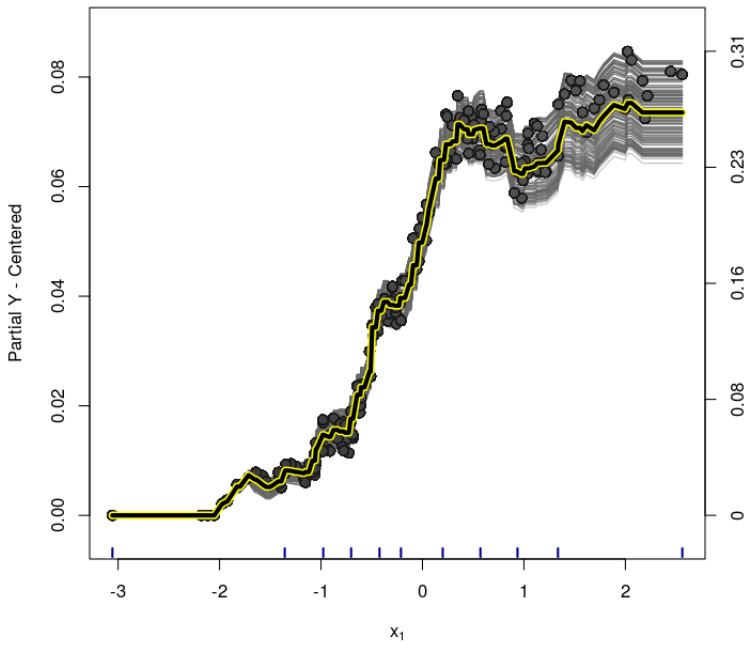

Figure B.29: *Oracle model (Toy Example) - Variable x_1 - centered-ICE Plot for the treatment effect.*

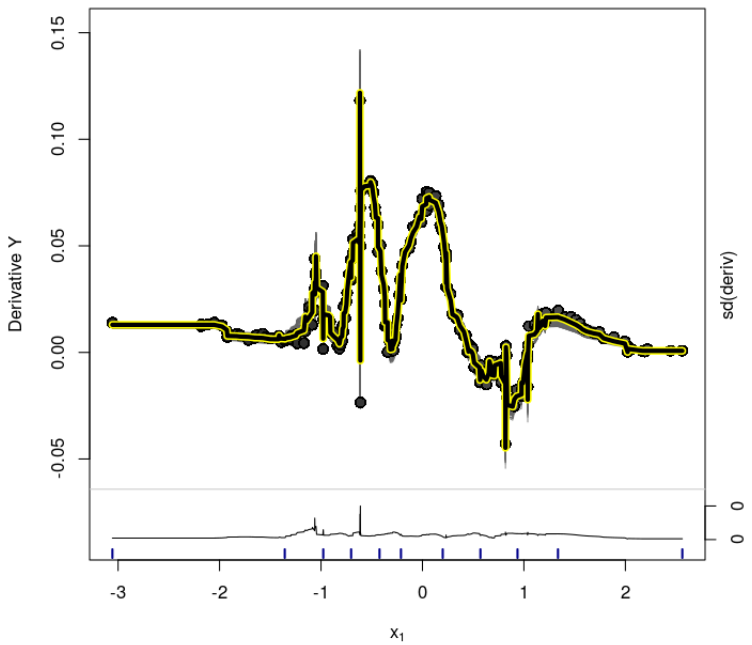

Figure B.30: *Oracle model (Toy Example) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

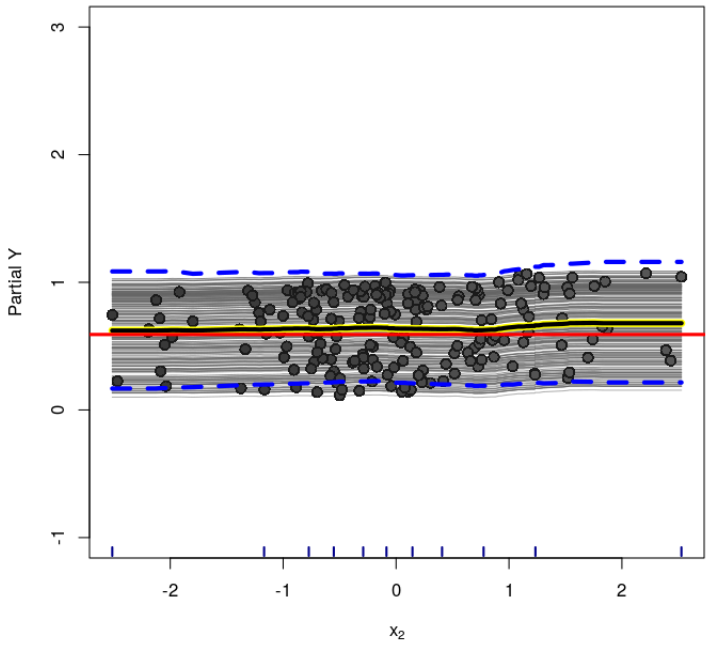

Figure B.31: *Oracle model (Toy Example) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

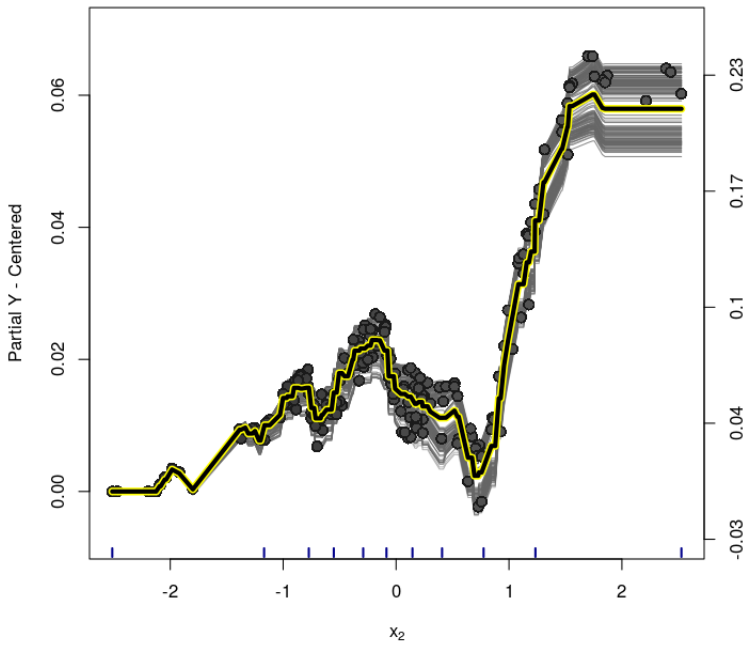

Figure B.32: *Oracle model (Toy Example) - Variable x_2 - centered-ICE Plot for the treatment effect.*

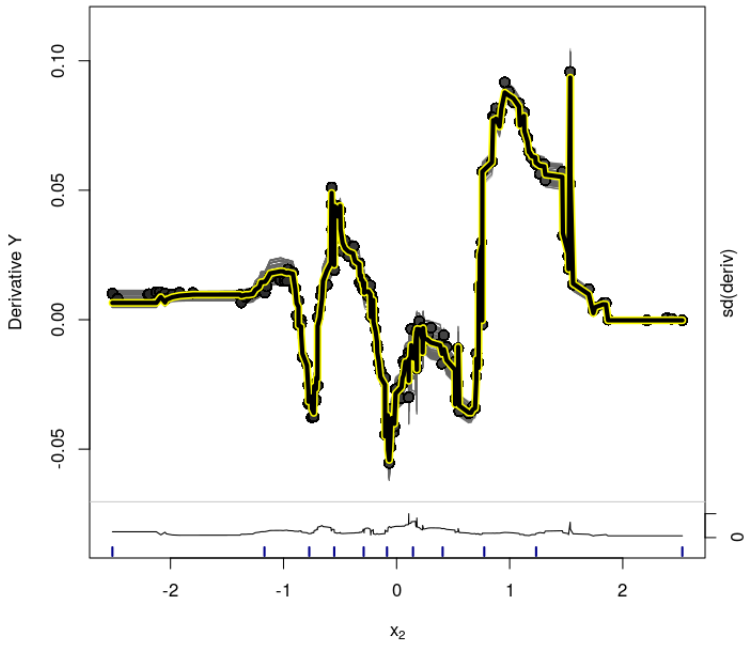

Figure B.33: *Oracle model (Toy Example) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

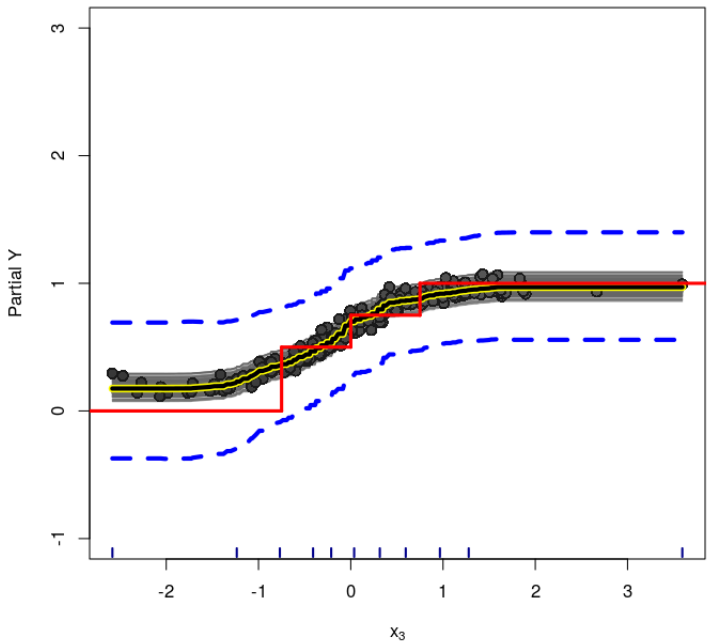

Figure B.34: *Oracle model (Toy Example) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

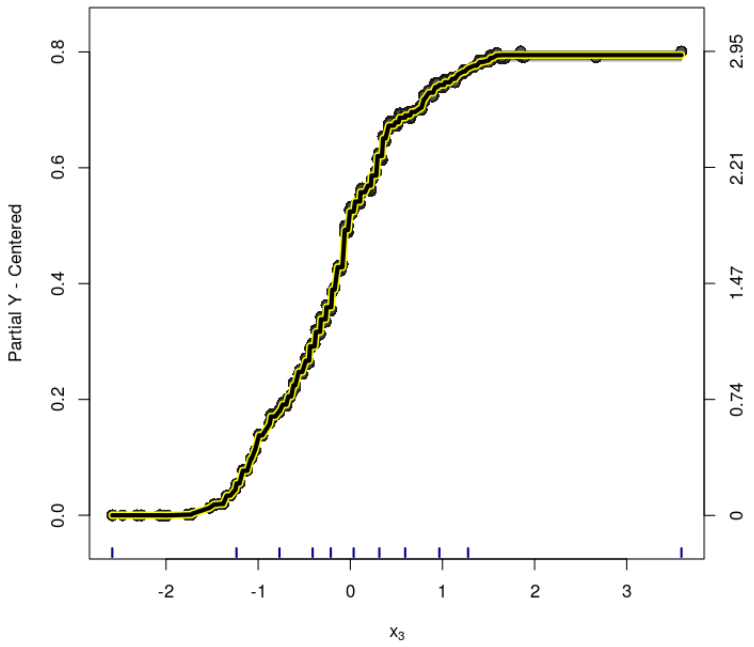

Figure B.35: *Oracle model (Toy Example) - Variable x_3 - centered-ICE Plot for the treatment effect.*

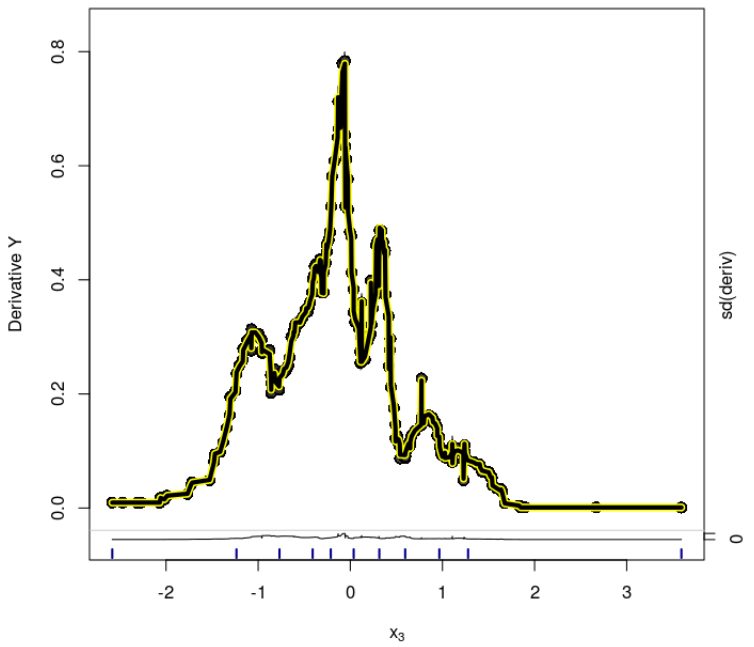

Figure B.36: *Oracle model (Toy Example) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

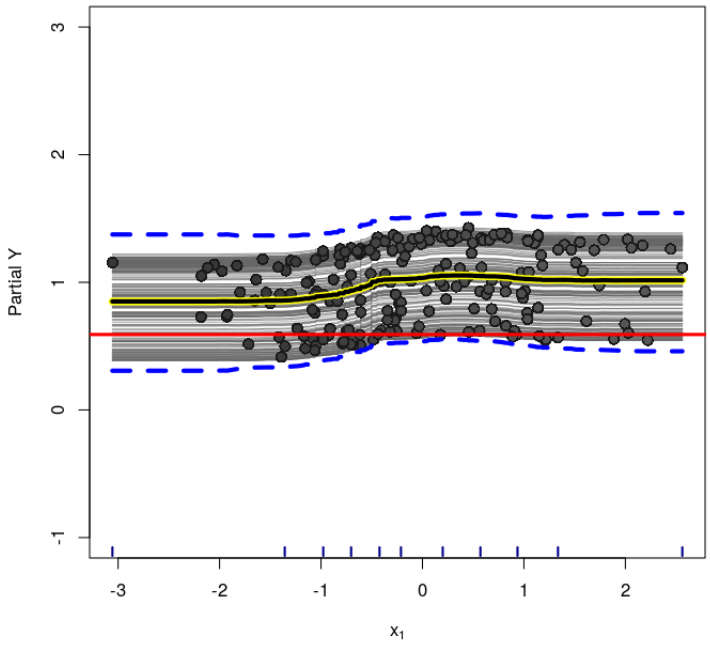

Figure B.37: *PS-BART model (Toy Example) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

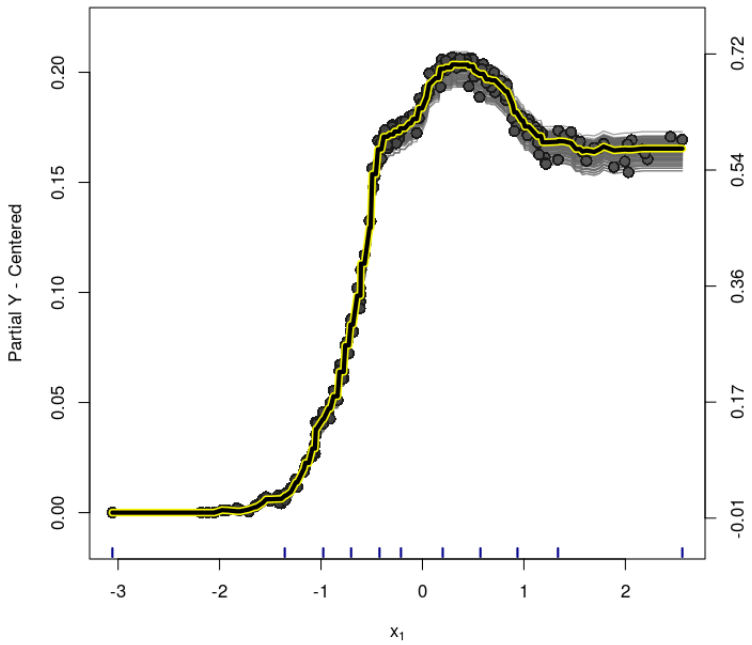

Figure B.38: *PS-BART model (Toy Example) - Variable x_1 - centered-ICE Plot for the treatment effect.*

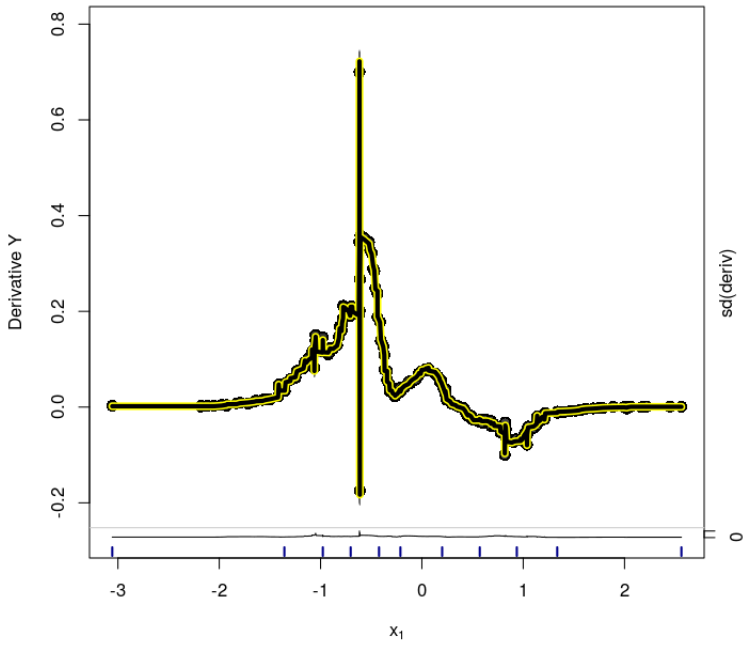

Figure B.39: *PS-BART model (Toy Example) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

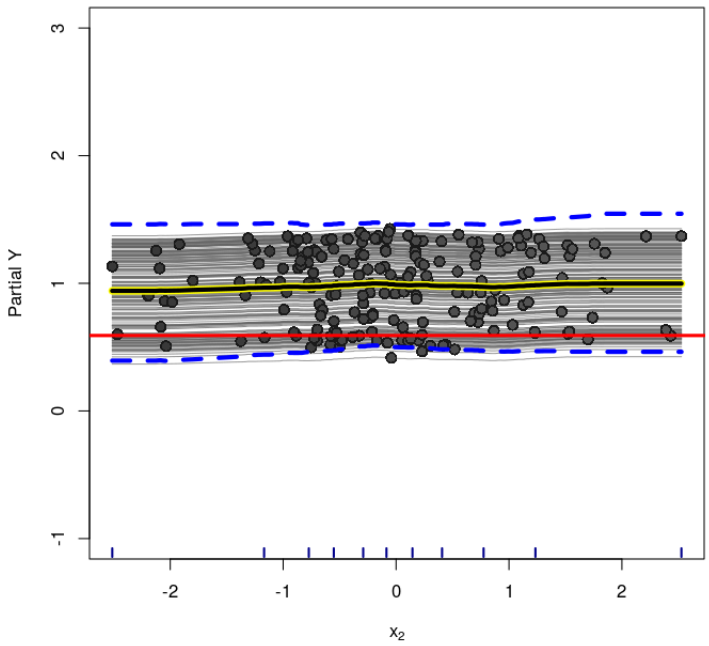

Figure B.40: *PS-BART model (Toy Example) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

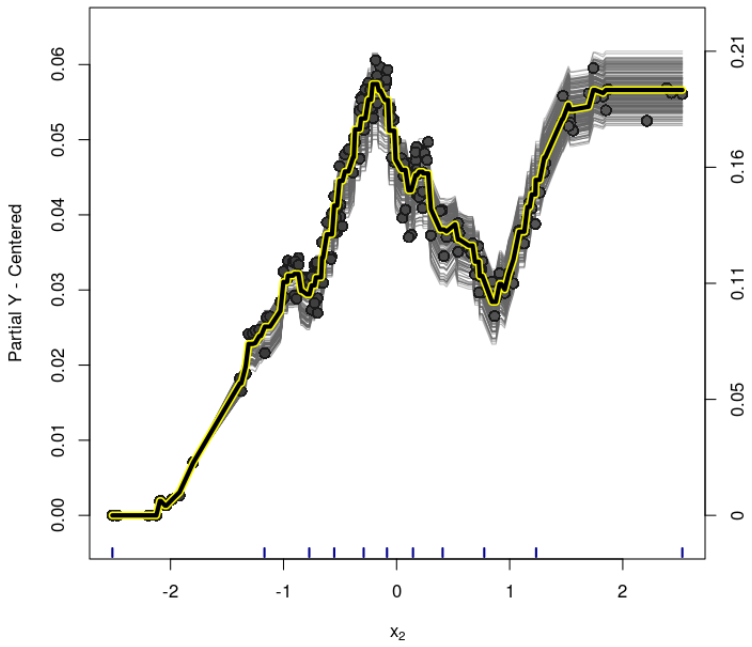

Figure B.41: *PS-BART model (Toy Example) - Variable x_2 - centered-ICE Plot for the treatment effect.*

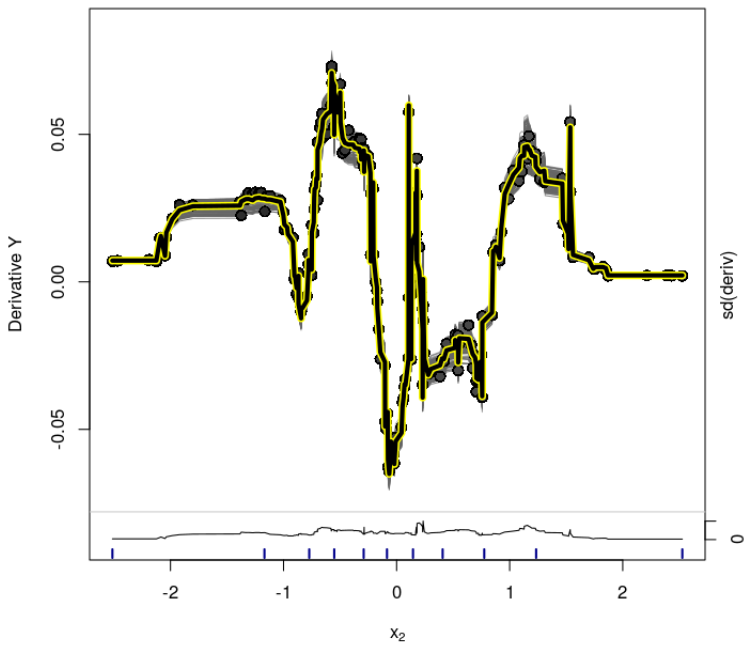

Figure B.42: *PS-BART model (Toy Example) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

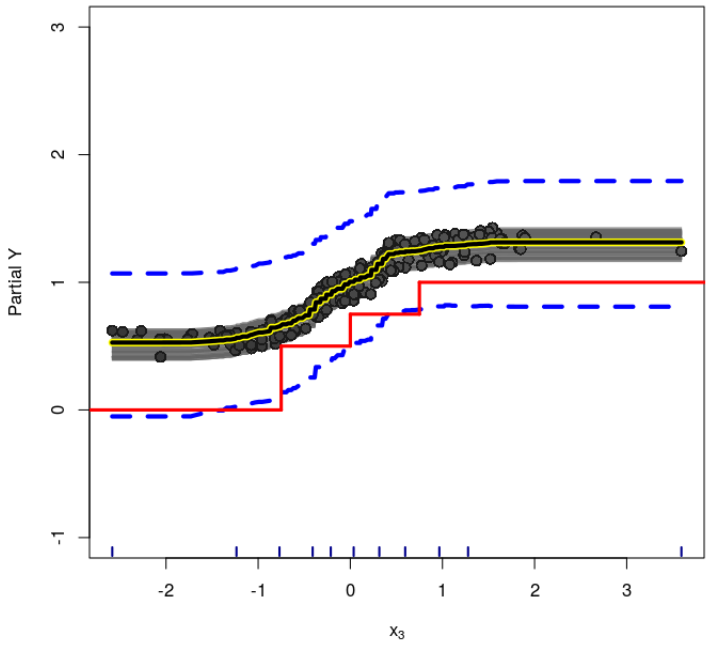

Figure B.43: *PS-BART model (Toy Example) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

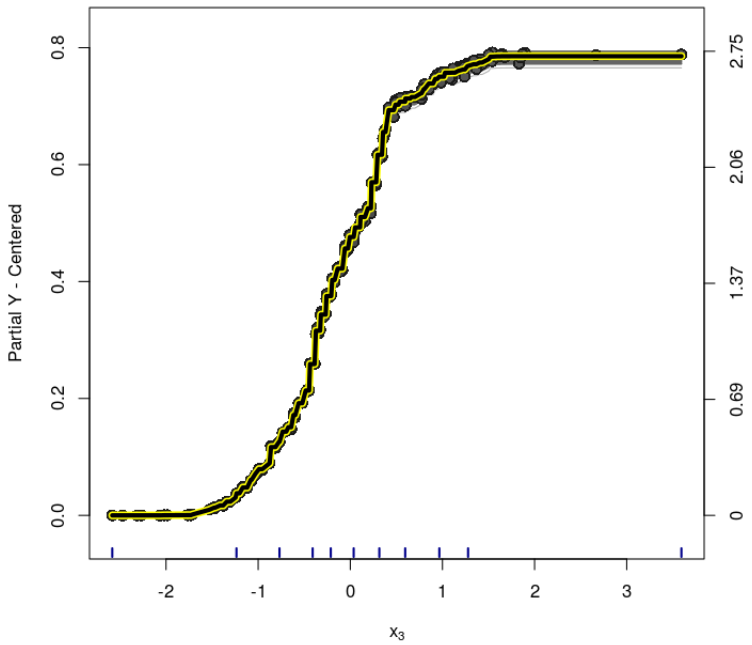

Figure B.44: *PS-BART model (Toy Example) - Variable x_3 - centered-ICE Plot for the treatment effect.*

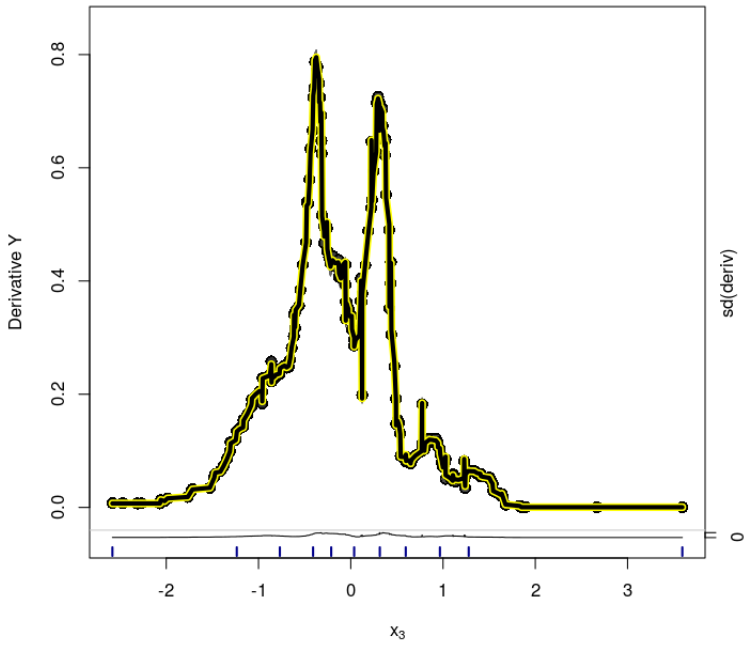

Figure B.45: *PS-BART model (Toy Example) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

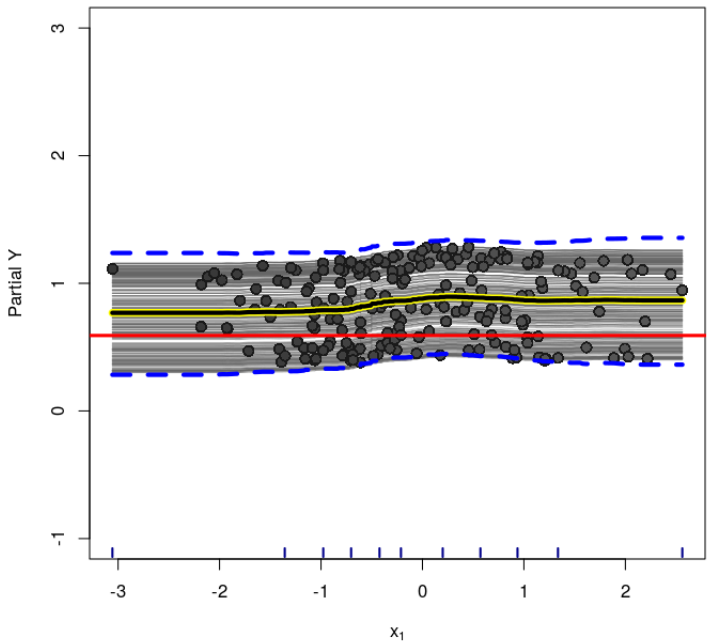

Figure B.46: *GLM-BART model (Toy Example) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

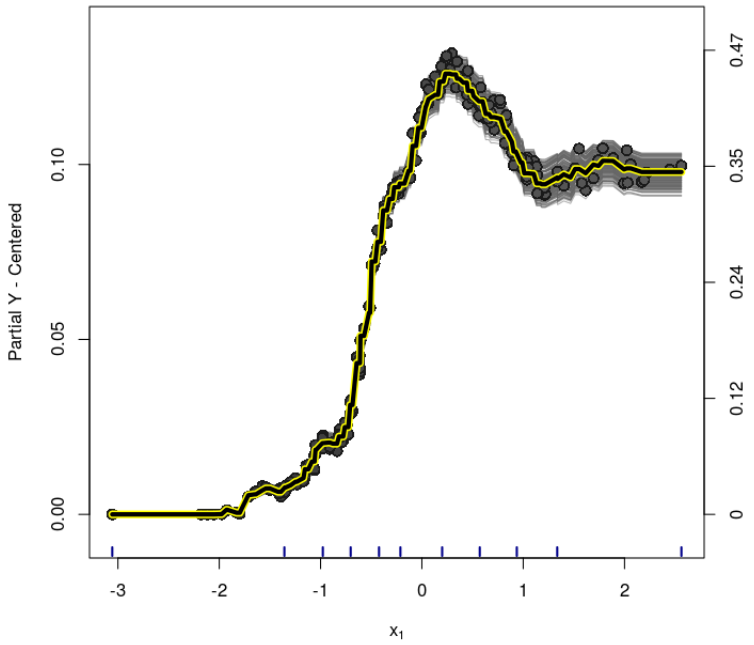

Figure B.47: *GLM-BART model (Toy Example) - Variable x_1 - centered-ICE Plot for the treatment effect.*

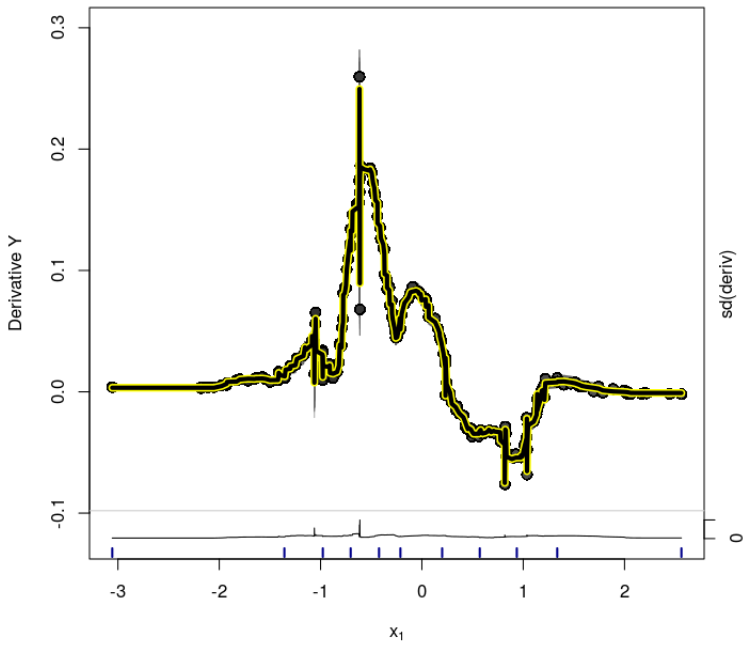

Figure B.48: *GLM-BART model (Toy Example) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

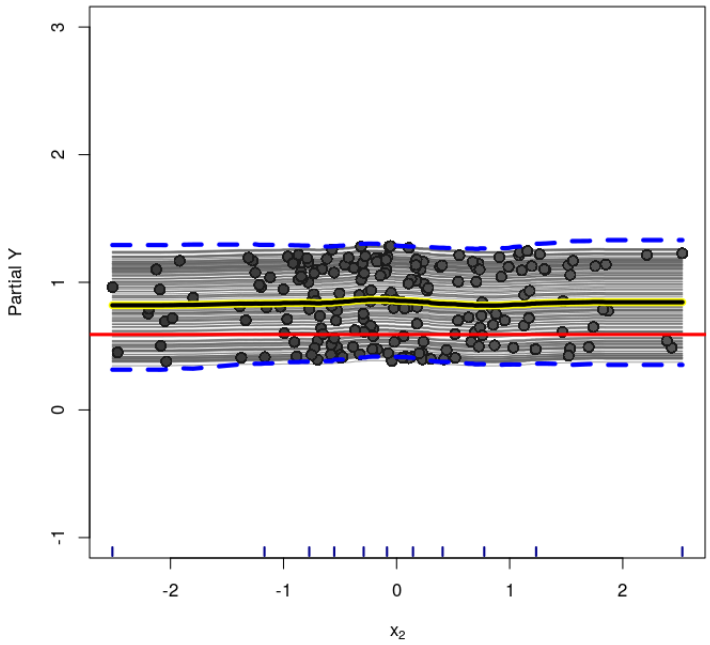

Figure B.49: *GLM-BART model (Toy Example) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

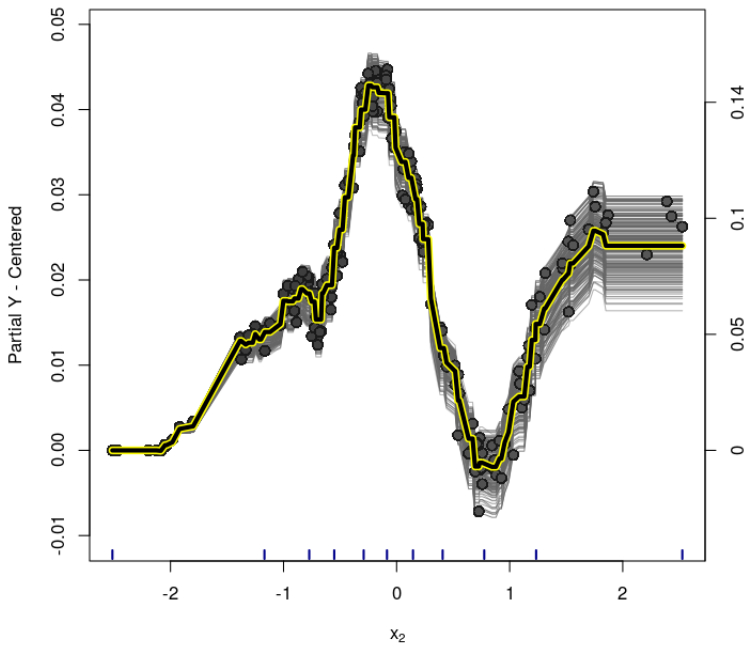

Figure B.50: *GLM-BART model (Toy Example) - Variable x_2 - centered-ICE Plot for the treatment effect.*

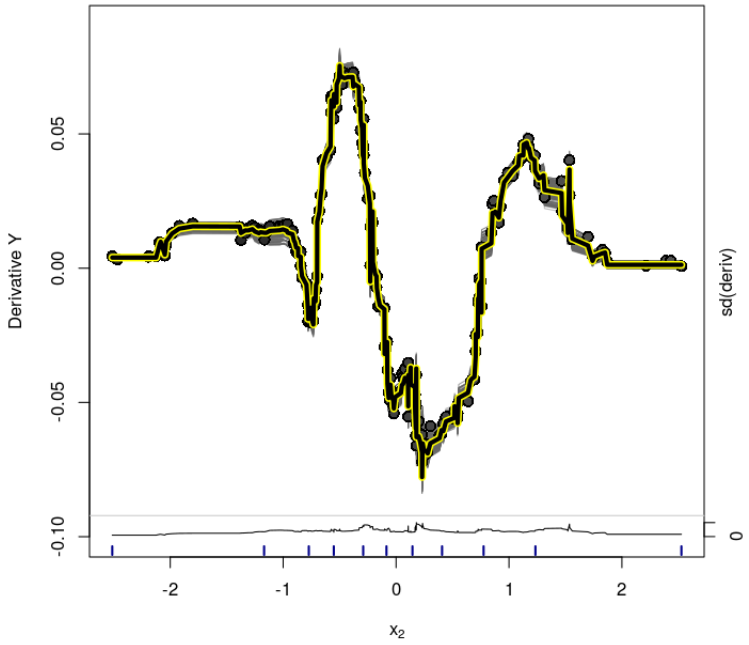

Figure B.51: *GLM-BART model (Toy Example) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

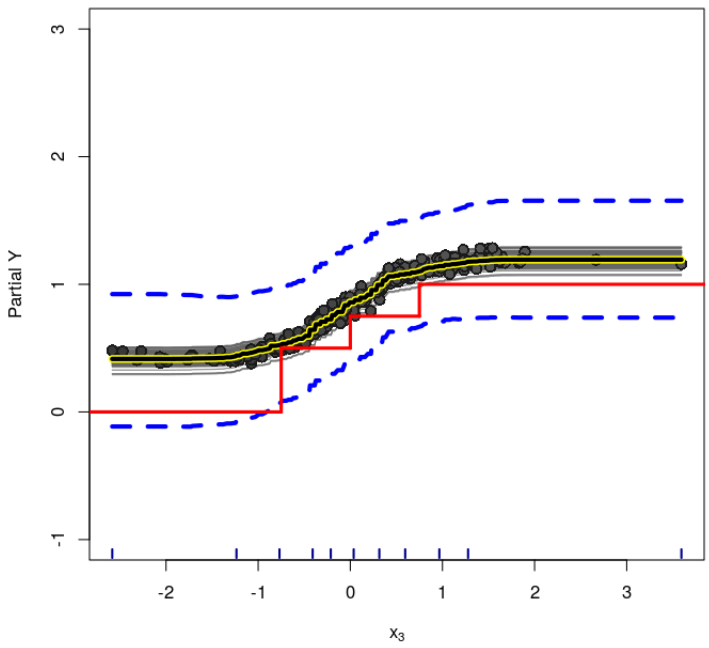

Figure B.52: *GLM-BART model (Toy Example) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

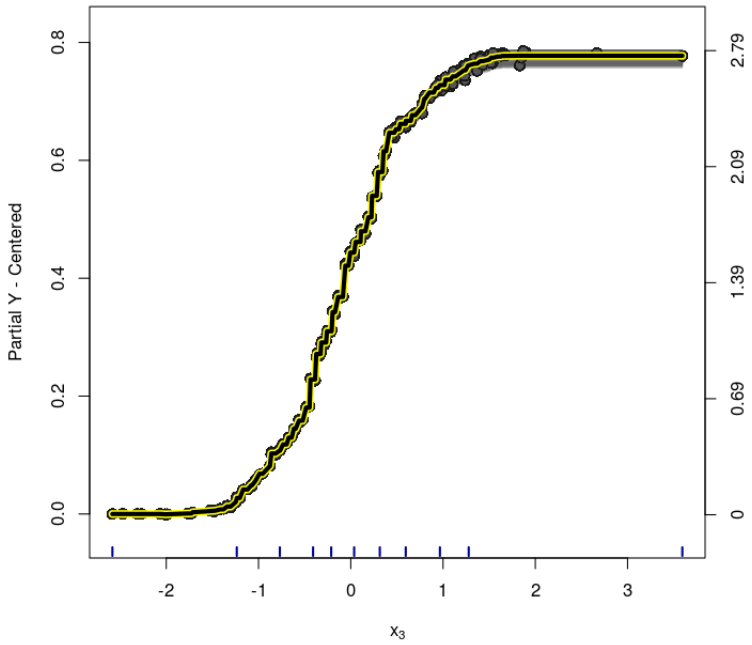

Figure B.53: *GLM-BART model (Toy Example) - Variable x_3 - centered-ICE Plot for the treatment effect.*

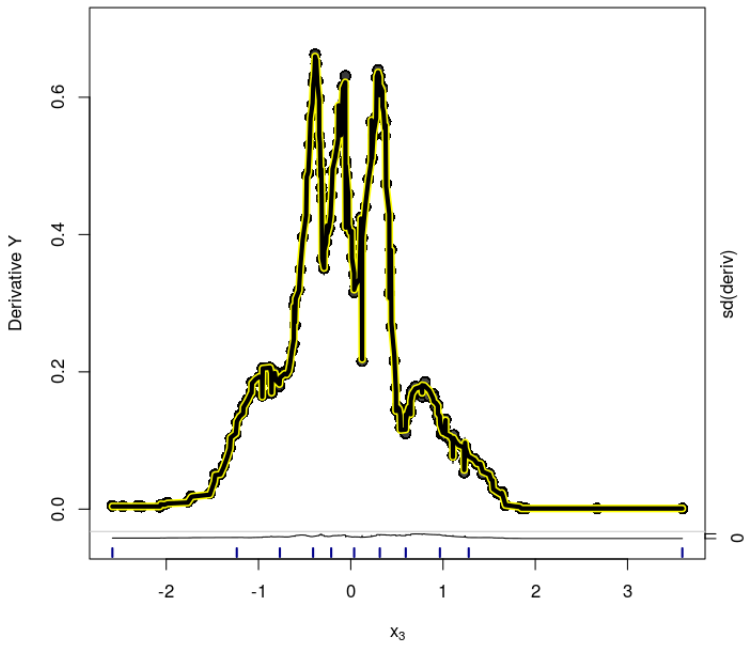

Figure B.54: *GLM-BART model (Toy Example) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

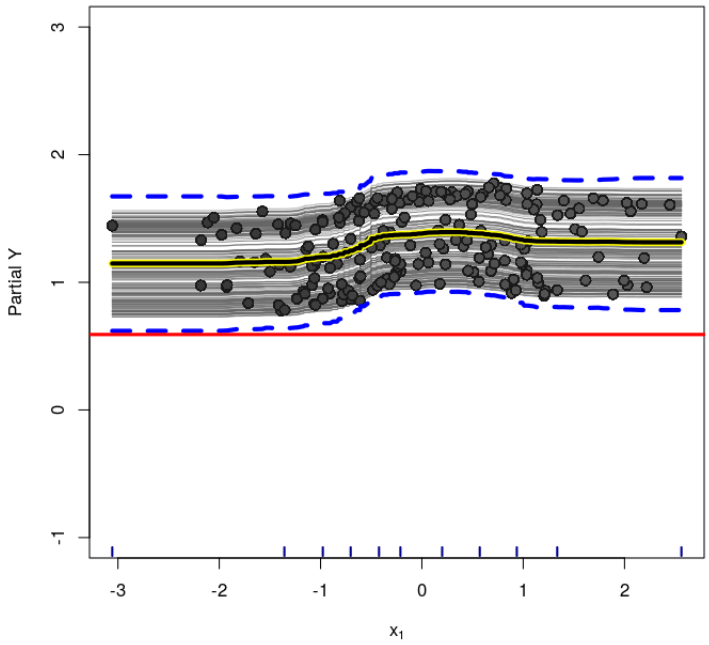

Figure B.55: *Rand-BART model (Toy Example) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

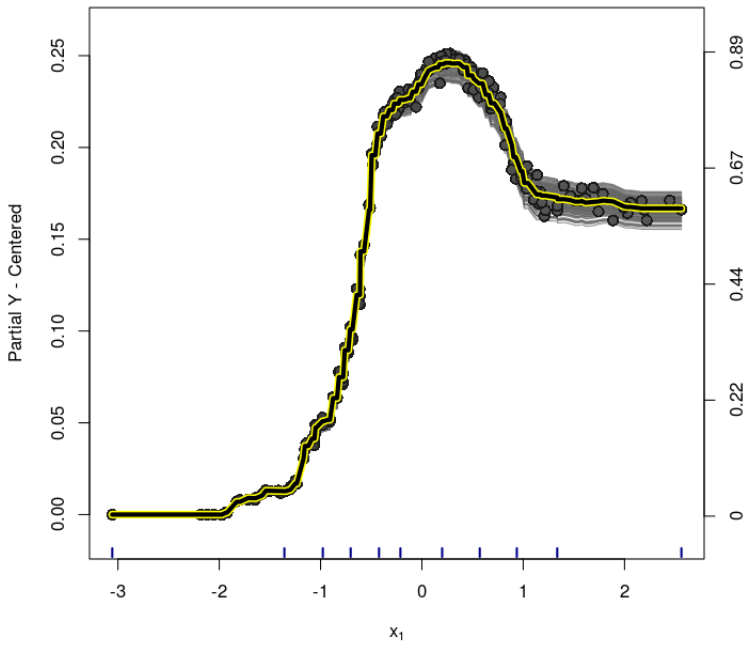

Figure B.56: *Rand-BART model (Toy Example) - Variable x_1 - centered-ICE Plot for the treatment effect.*

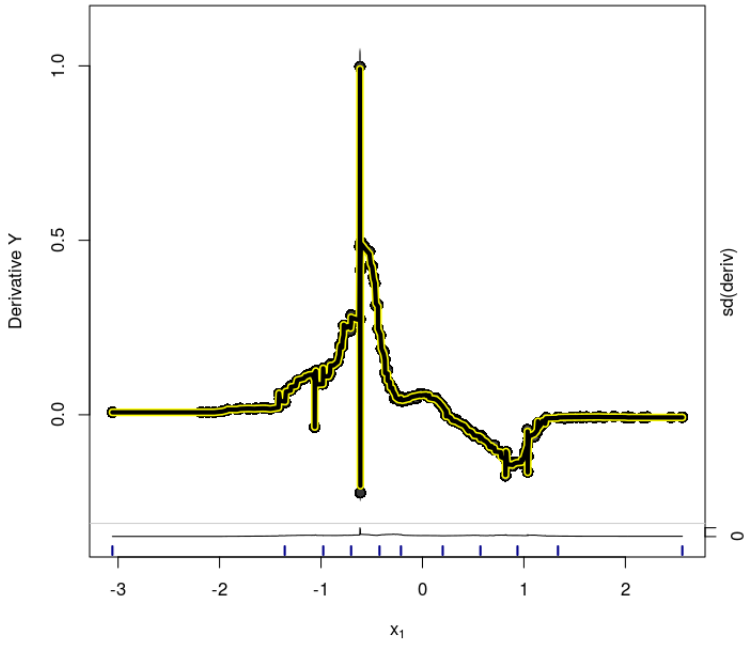

Figure B.57: *Rand-BART model (Toy Example) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

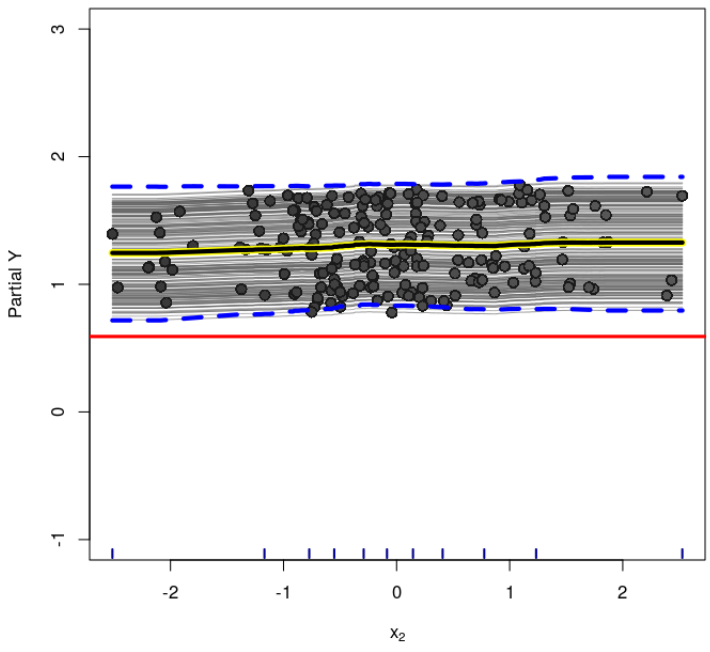

Figure B.58: *Rand-BART model (Toy Example) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

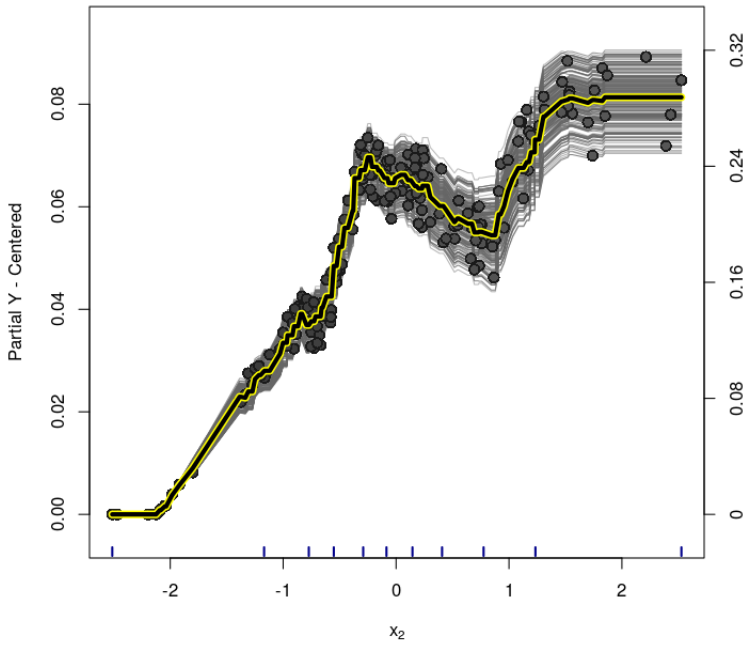

Figure B.59: *Rand-BART model (Toy Example) - Variable x_2 - centered-ICE Plot for the treatment effect.*

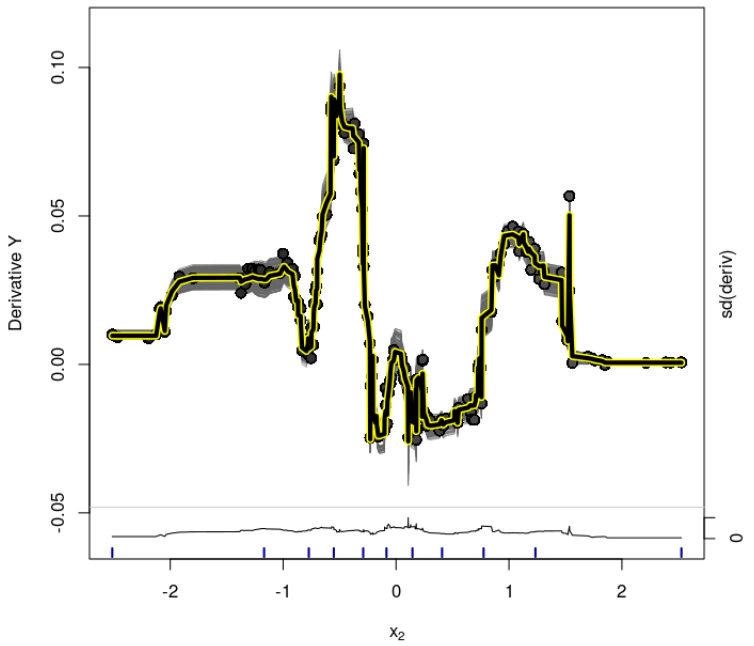

Figure B.60: *Rand-BART model (Toy Example) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

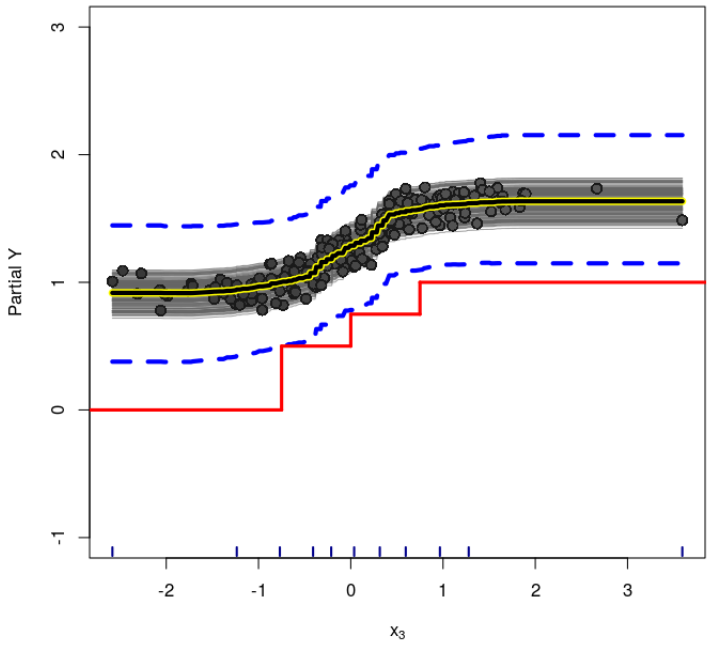

Figure B.61: *Rand-BART model (Toy Example) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

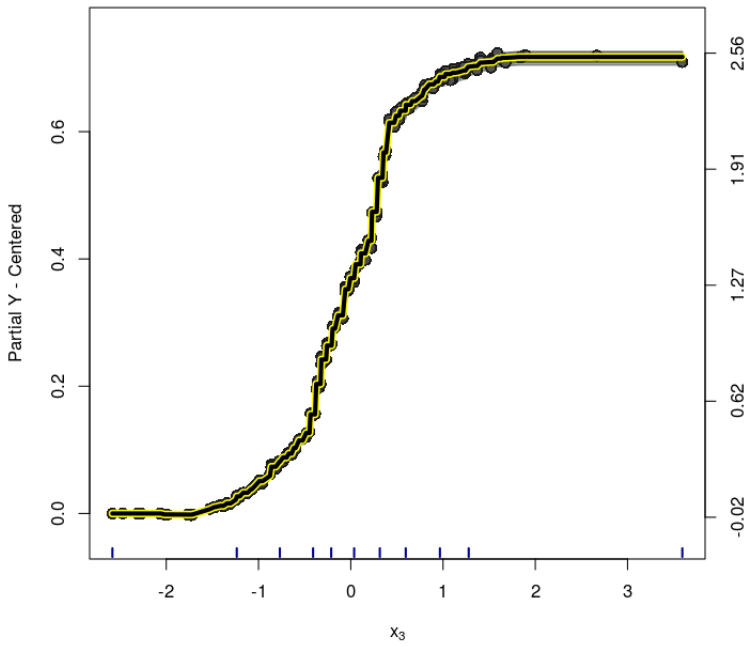

Figure B.62: *Rand-BART model (Toy Example) - Variable x_3 - centered-ICE Plot for the treatment effect.*

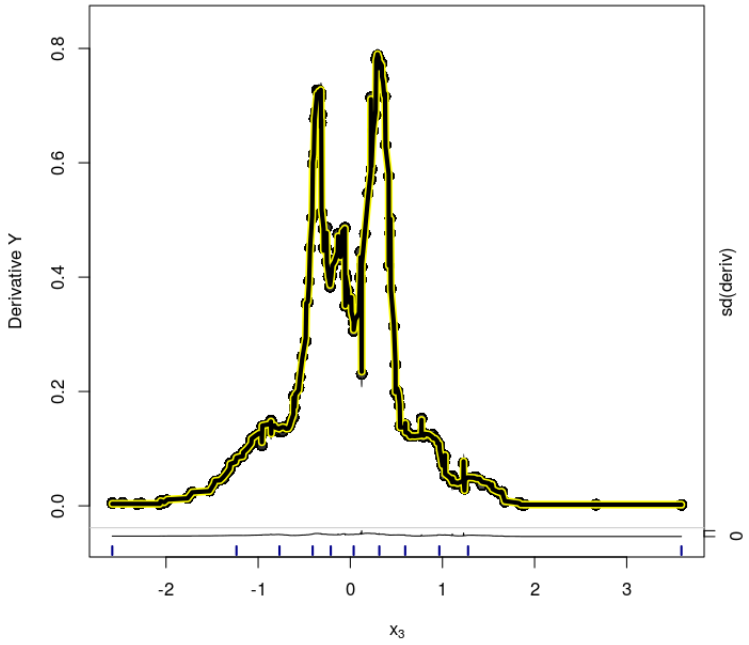

Figure B.63: *Rand-BART model (Toy Example) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

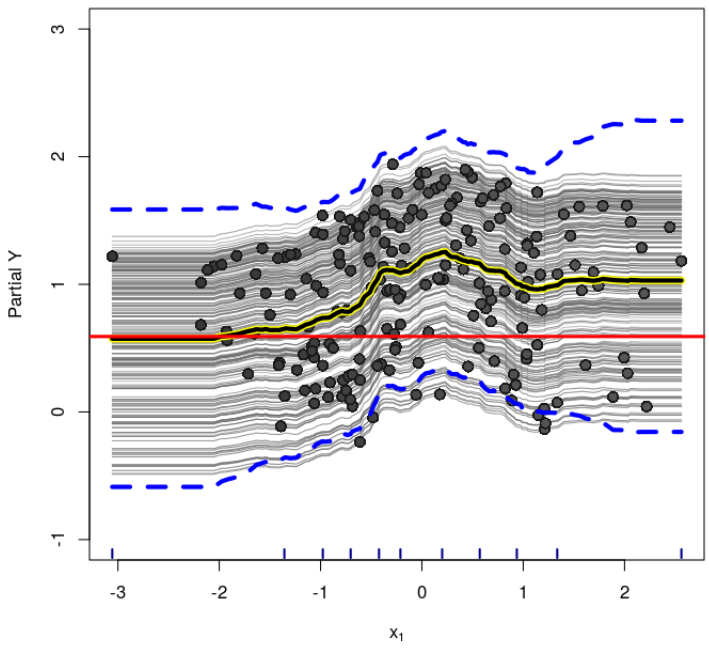

Figure B.64: *PS-BCF model (Toy Example) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

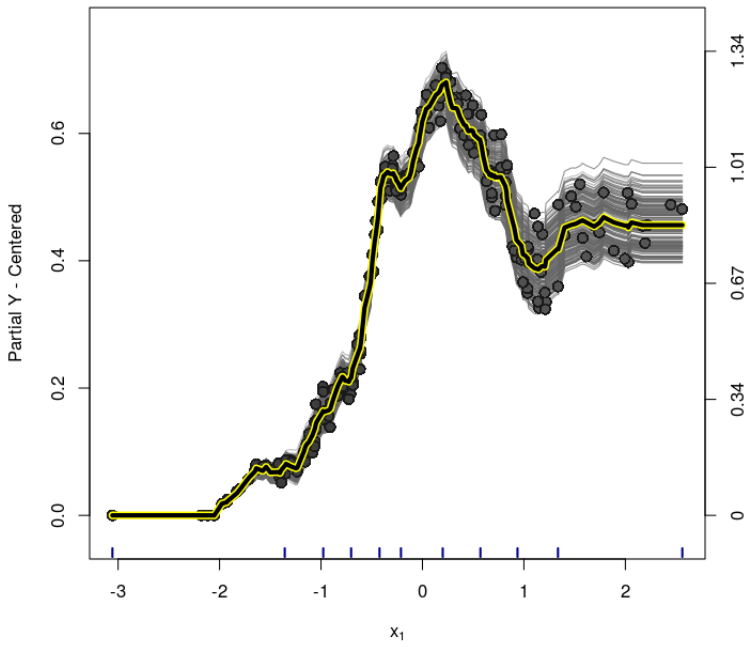

Figure B.65: *PS-BCF model (Toy Example) - Variable x_1 - centered-ICE Plot for the treatment effect.*

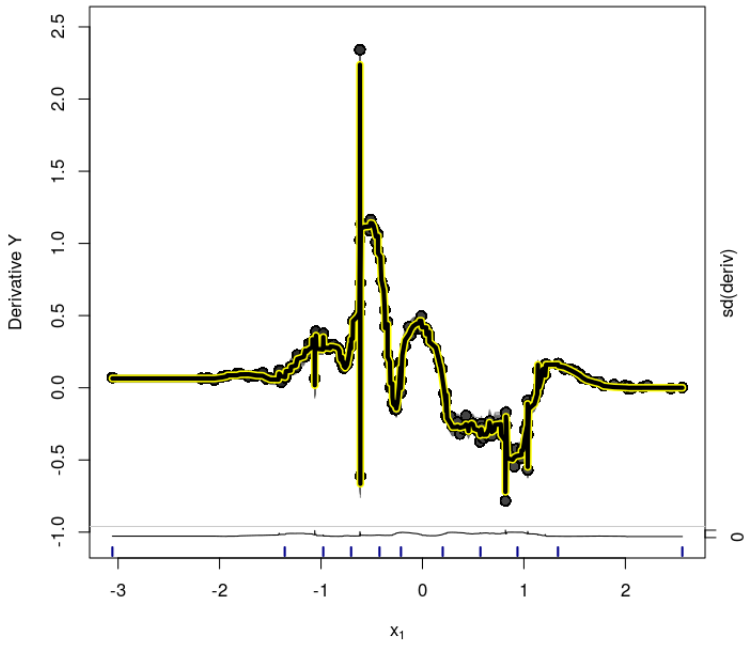

Figure B.66: *PS-BCF model (Toy Example) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

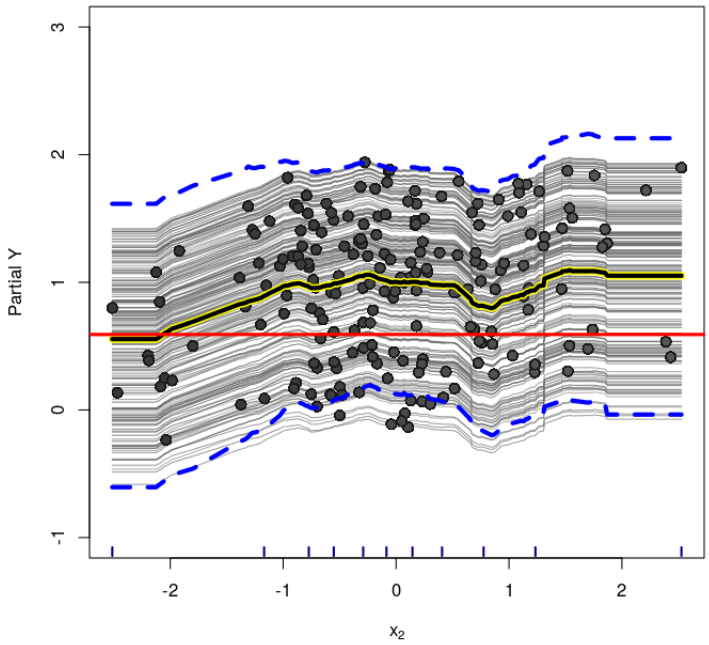

Figure B.67: *PS-BCF model (Toy Example) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

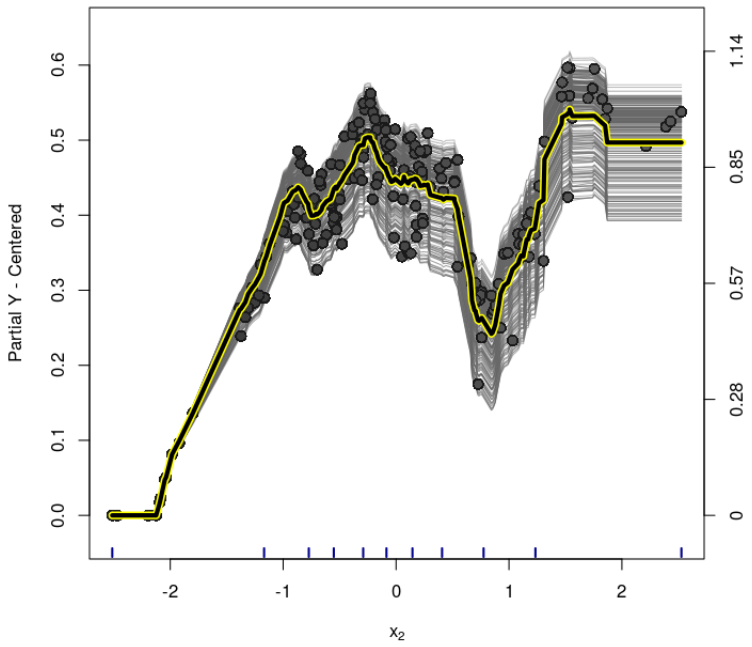

Figure B.68: *PS-BCF model (Toy Example) - Variable x_2 - centered-ICE Plot for the treatment effect.*

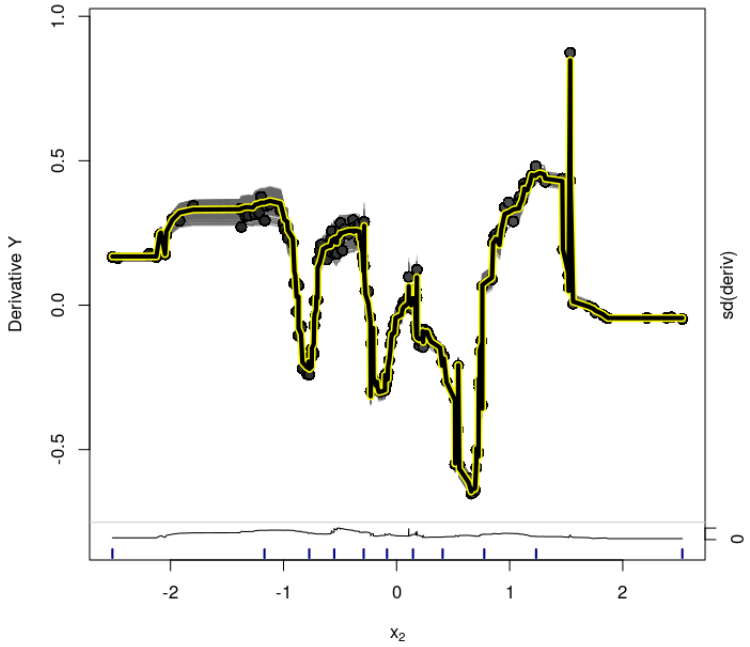

Figure B.69: *PS-BCF model (Toy Example) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

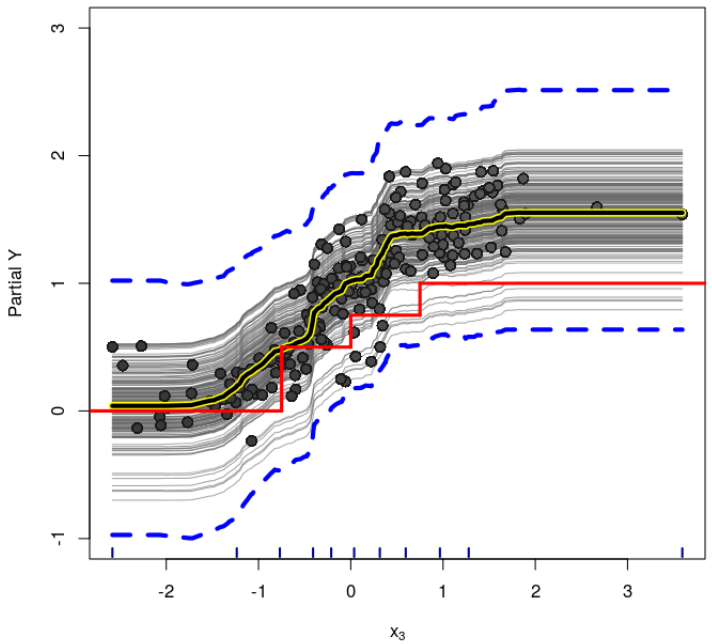

Figure B.70: *PS-BCF model (Toy Example) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

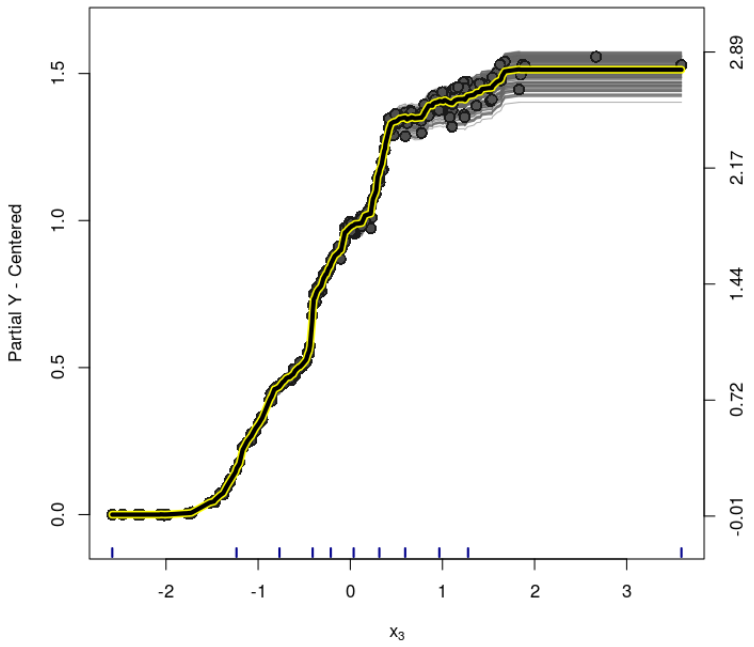

Figure B.71: *PS-BCF model (Toy Example) - Variable x_3 - centered-ICE Plot for the treatment effect.*

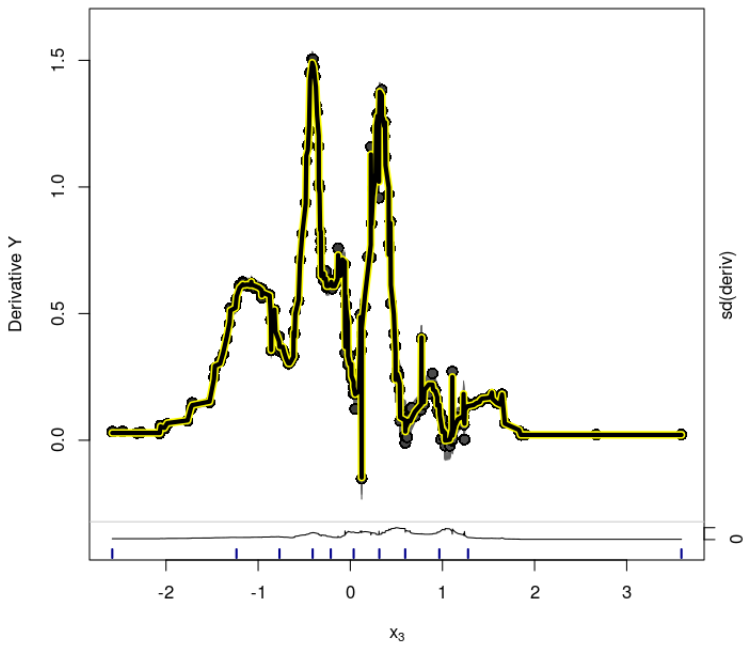

Figure B.72: *PS-BCF model (Toy Example) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

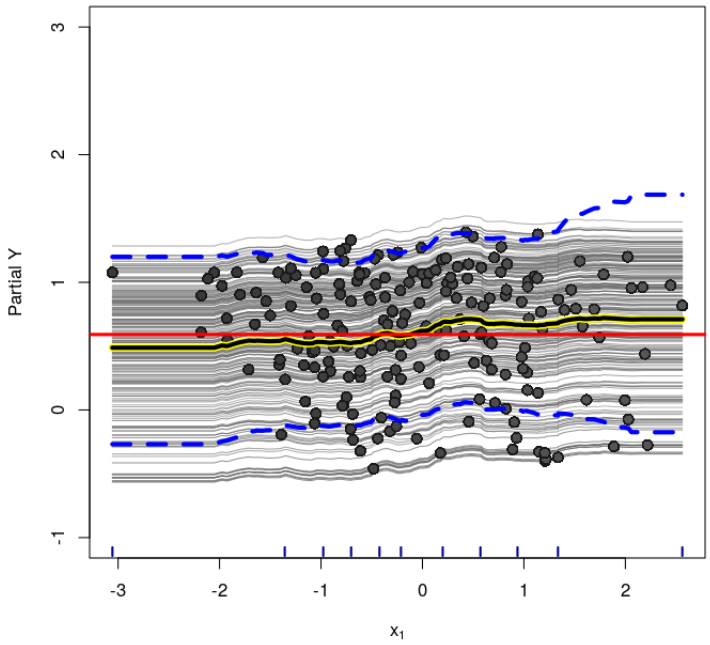

Figure B.73: *Oracle-BCF model (Toy Example) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

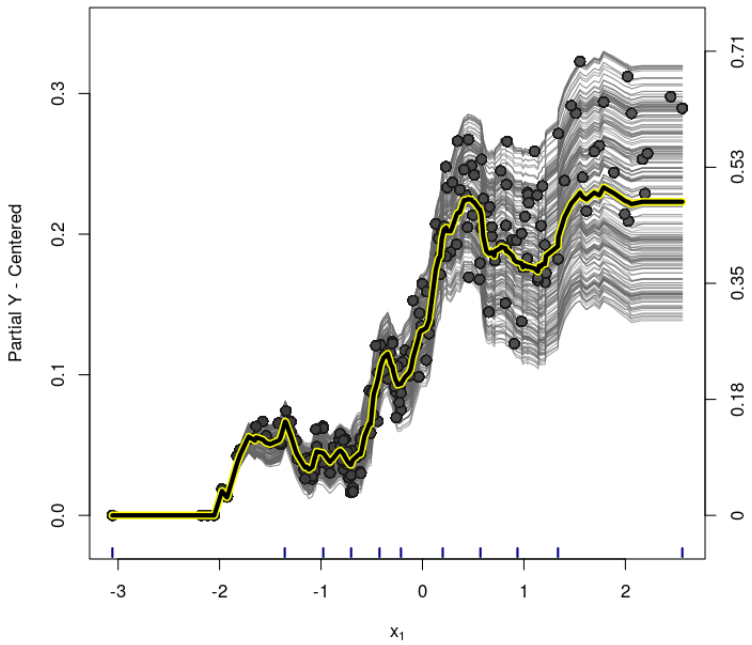

Figure B.74: *Oracle-BCF model (Toy Example) - Variable x_1 - centered-ICE Plot for the treatment effect.*

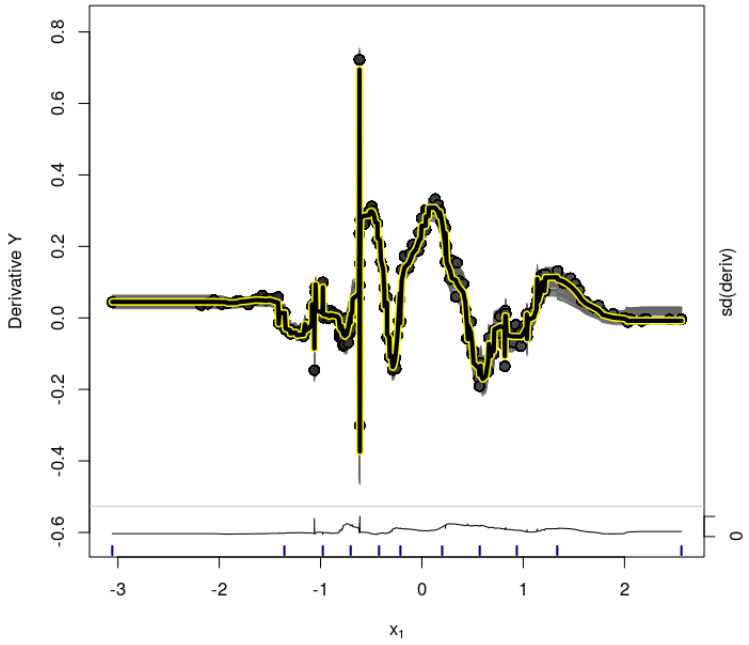

Figure B.75: *Oracle-BCF model (Toy Example) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

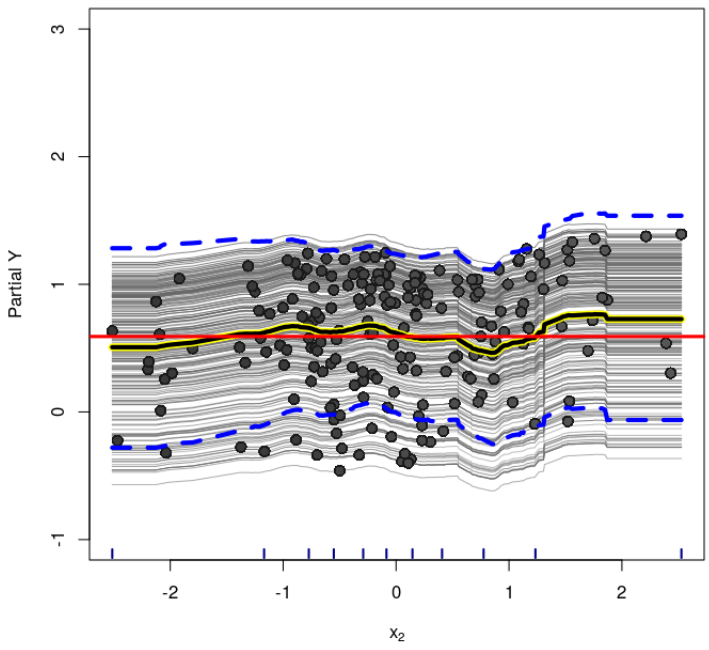

Figure B.76: *Oracle-BCF model (Toy Example) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

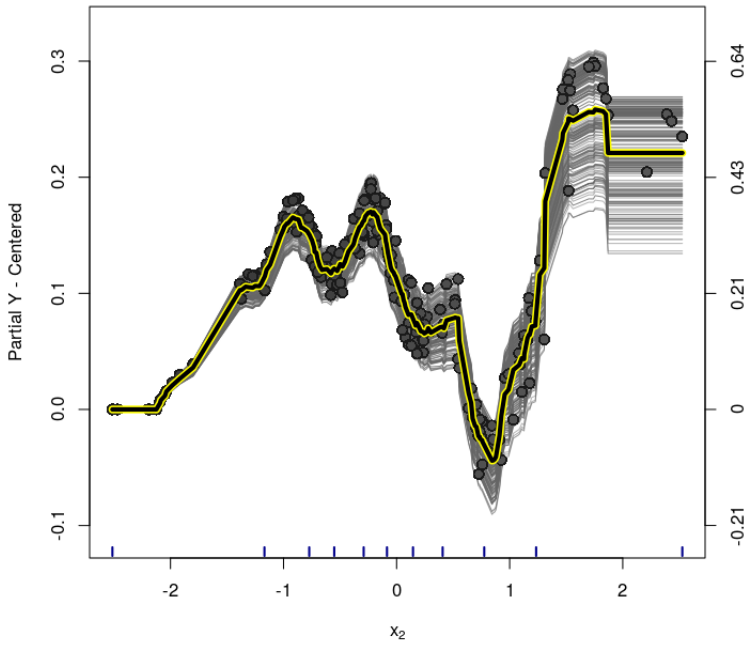

Figure B.77: *Oracle-BCF model (Toy Example) - Variable x_2 - centered-ICE Plot for the treatment effect.*

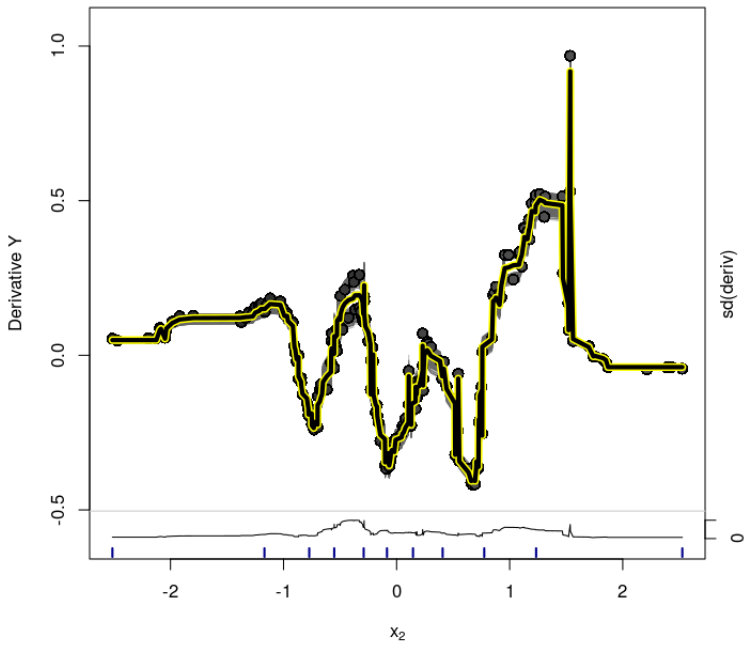

Figure B.78: *Oracle-BCF model (Toy Example) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

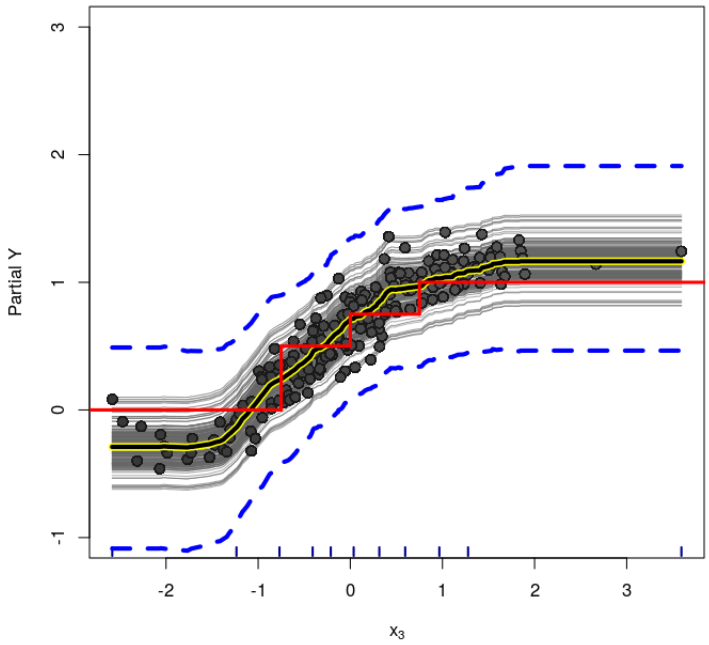

Figure B.79: *Oracle-BCF model (Toy Example) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

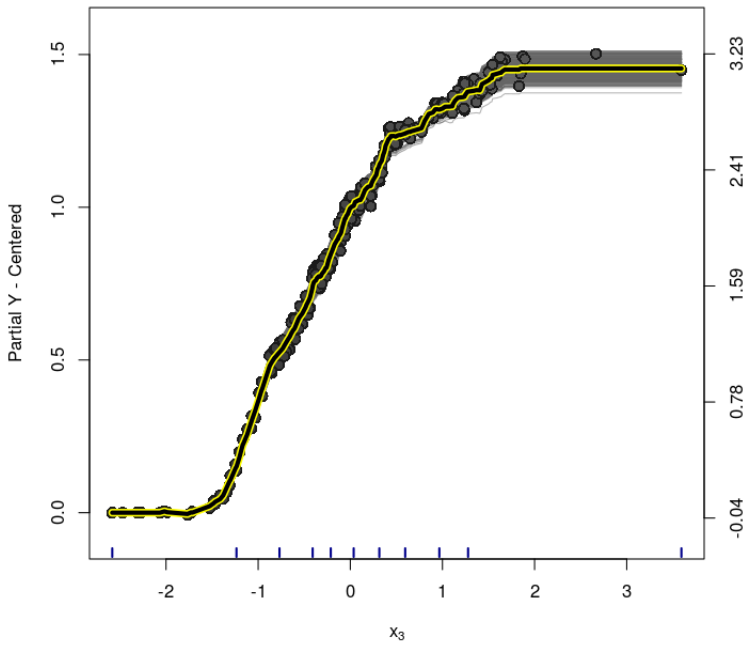

Figure B.80: *Oracle-BCF model (Toy Example) - Variable x_3 - centered-ICE Plot for the treatment effect.*

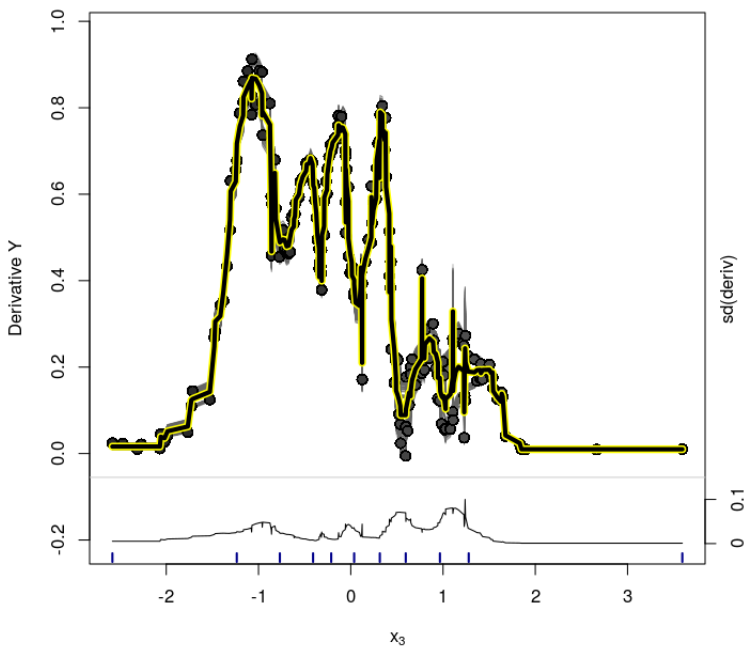

Figure B.81: *Oracle-BCF model (Toy Example) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

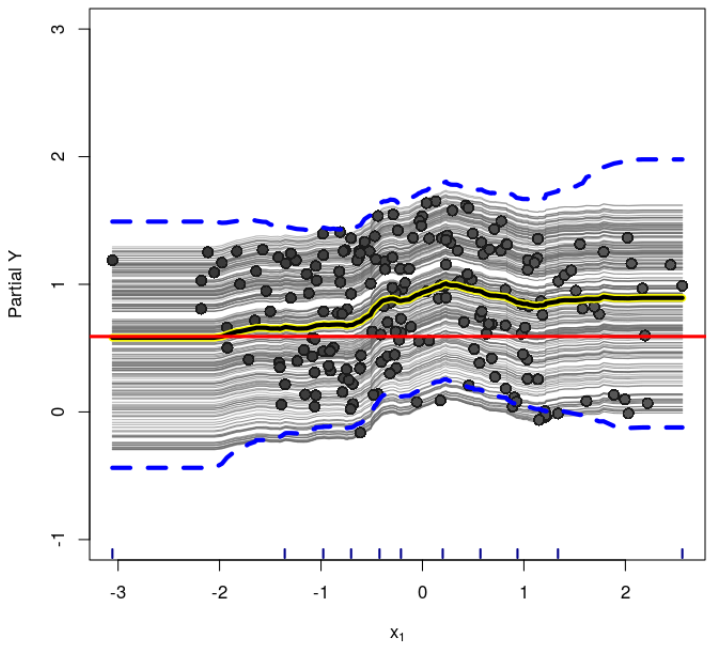

Figure B.82: *GLM-BCF model (Toy Example) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

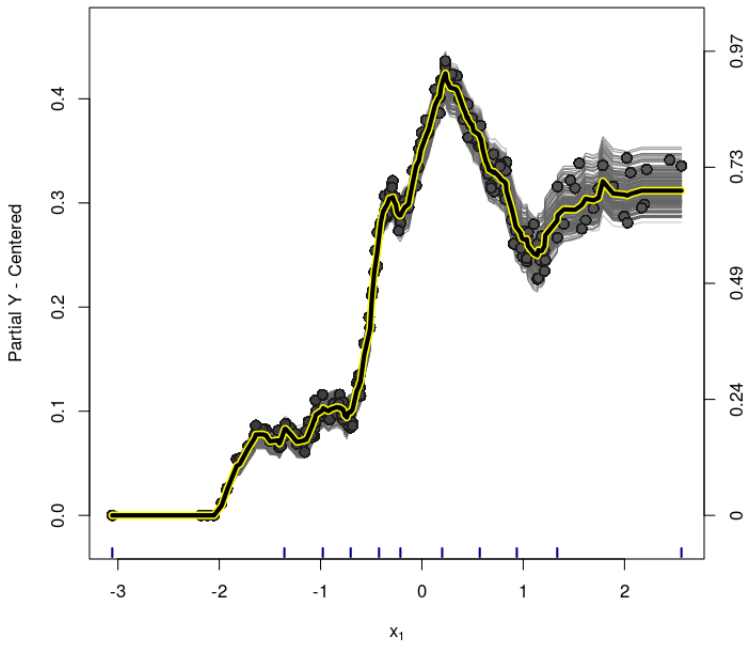

Figure B.83: *GLM-BCF model (Toy Example) - Variable x_1 - centered-ICE Plot for the treatment effect.*

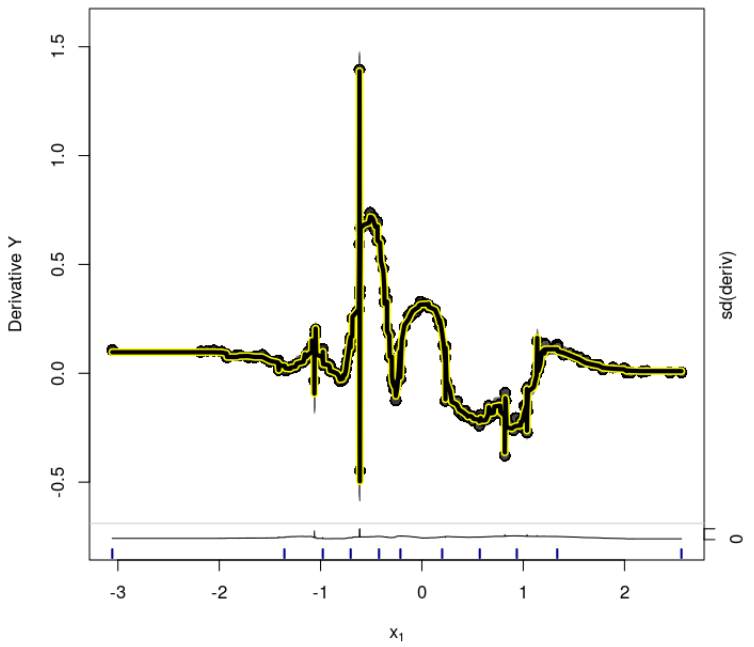

Figure B.84: *GLM-BCF model (Toy Example) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

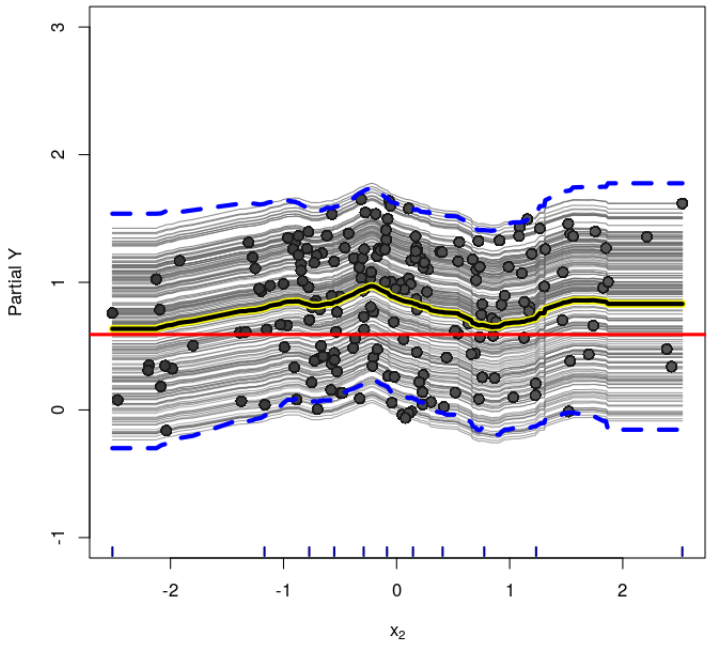

Figure B.85: *GLM-BCF model (Toy Example) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

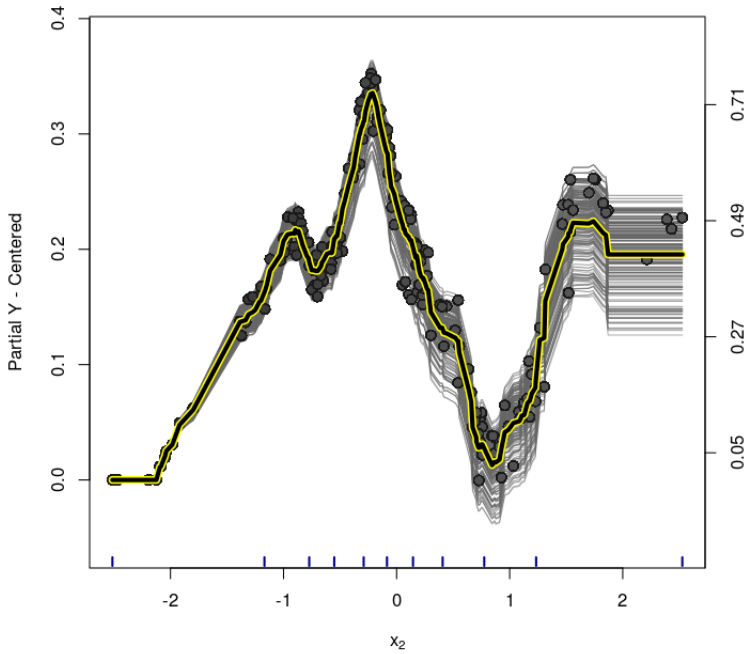

Figure B.86: *GLM-BCF model (Toy Example) - Variable x_2 - centered-ICE Plot for the treatment effect.*

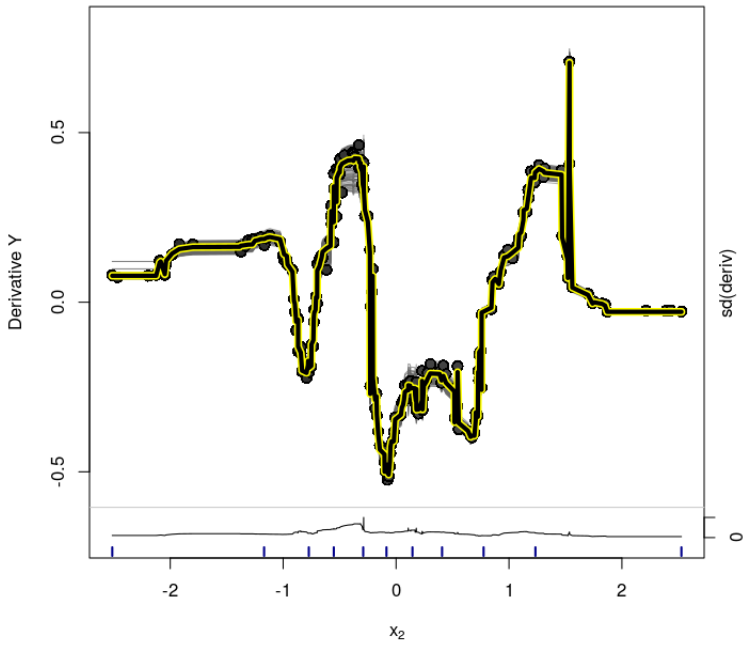

Figure B.87: *GLM-BCF model (Toy Example) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

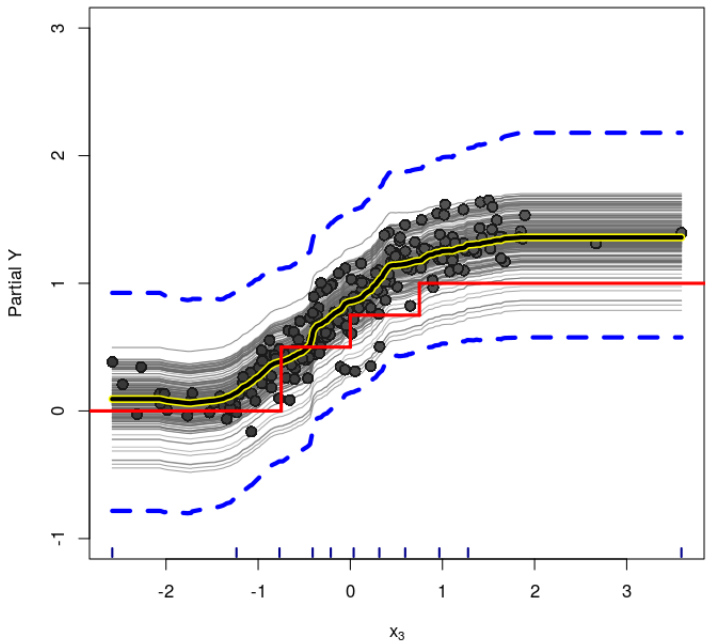

Figure B.88: *GLM-BCF model (Toy Example) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

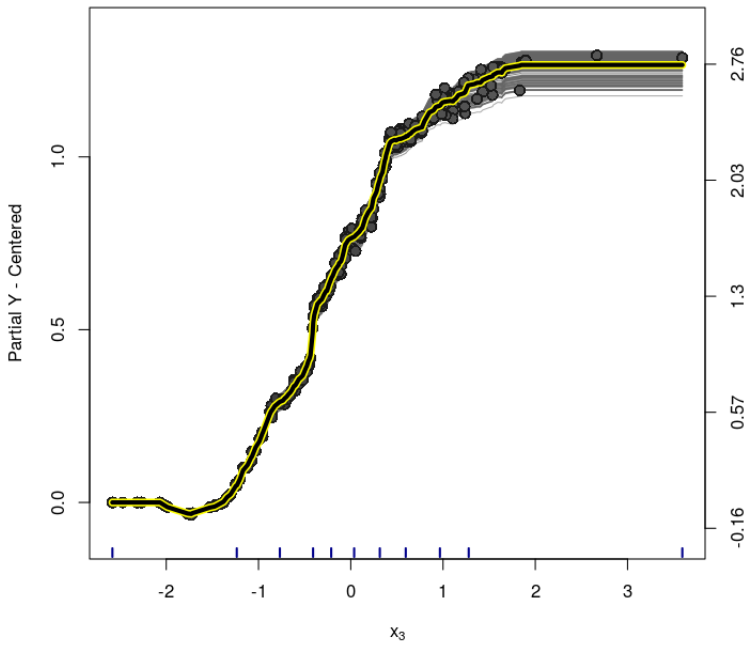

Figure B.89: *GLM-BCF model (Toy Example) - Variable x_3 - centered-ICE Plot for the treatment effect.*

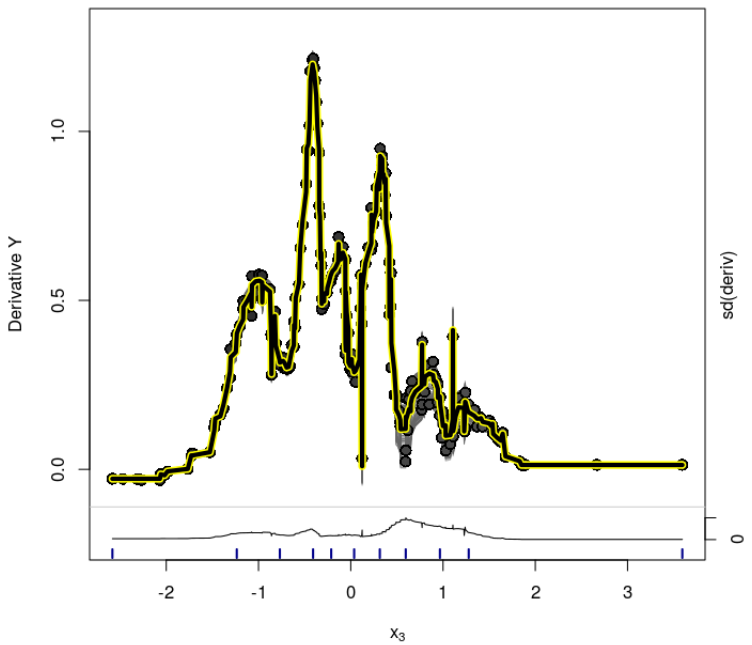

Figure B.90: *GLM-BCF model (Toy Example) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

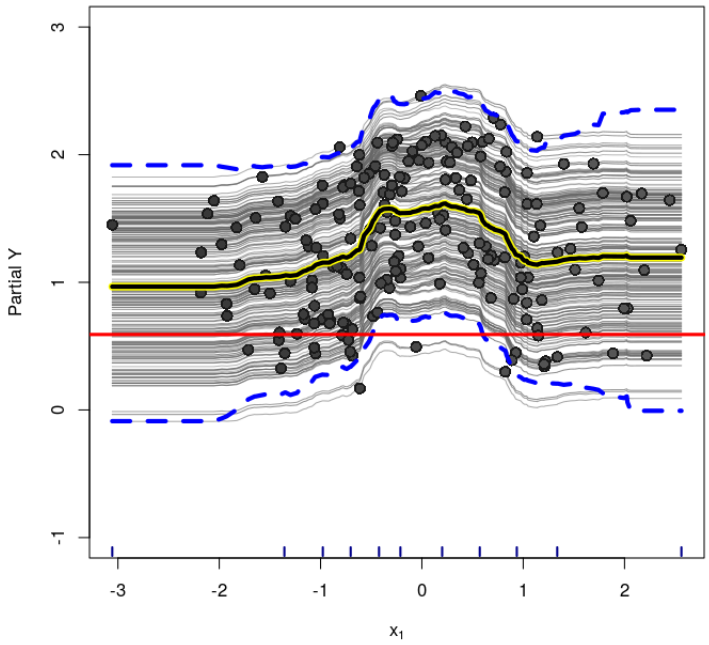

Figure B.91: *Rand-BCF model (Toy Example) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

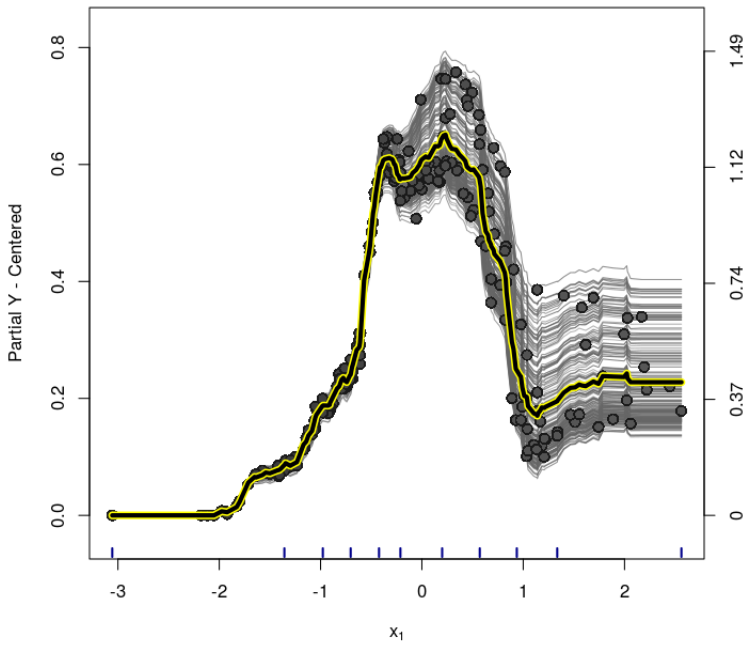

Figure B.92: *Rand-BCF model (Toy Example) - Variable x_1 - centered-ICE Plot for the treatment effect.*

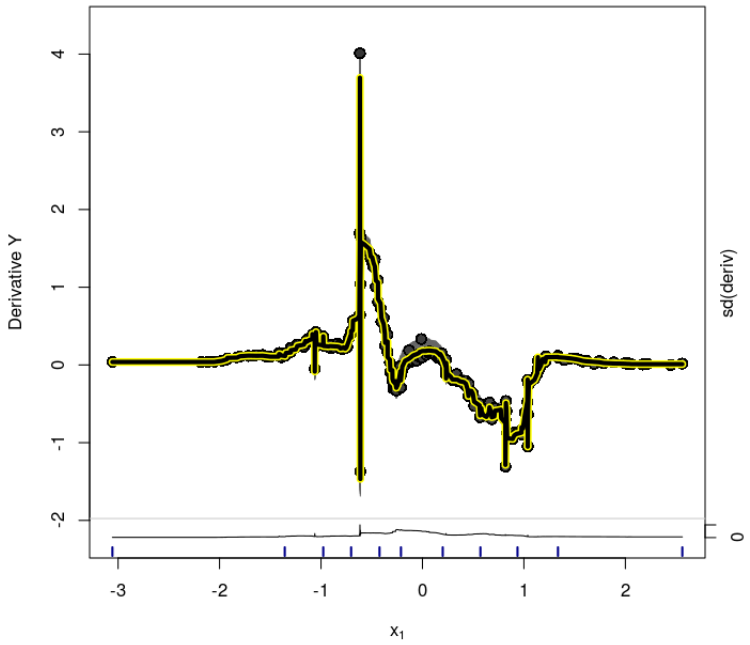

Figure B.93: *Rand-BCF model (Toy Example) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

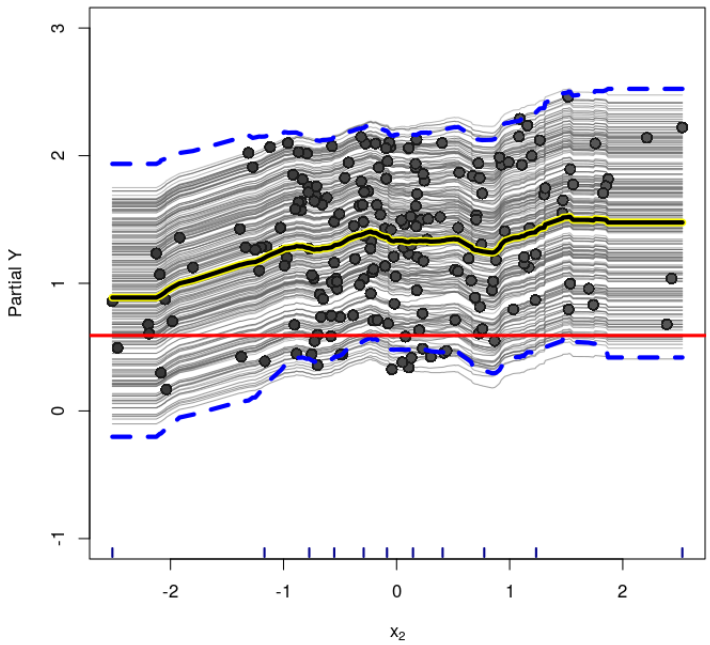

Figure B.94: *Rand-BCF model (Toy Example) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

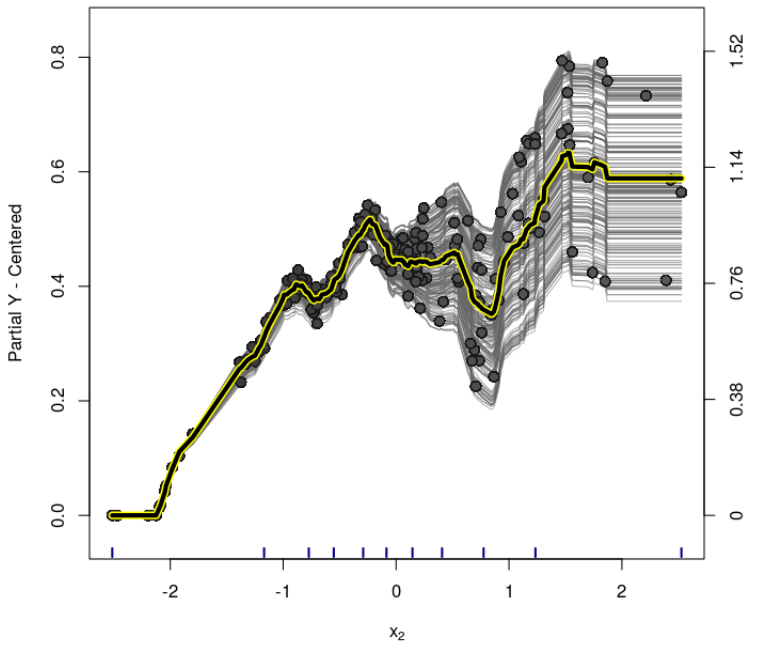

Figure B.95: *Rand-BCF model (Toy Example) - Variable x_2 - centered-ICE Plot for the treatment effect.*

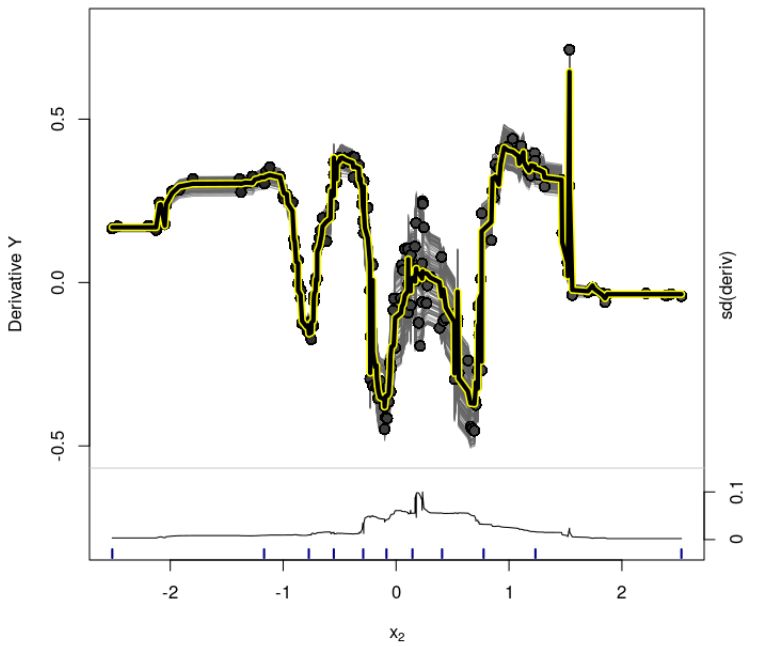

Figure B.96: *Rand-BCF model (Toy Example) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

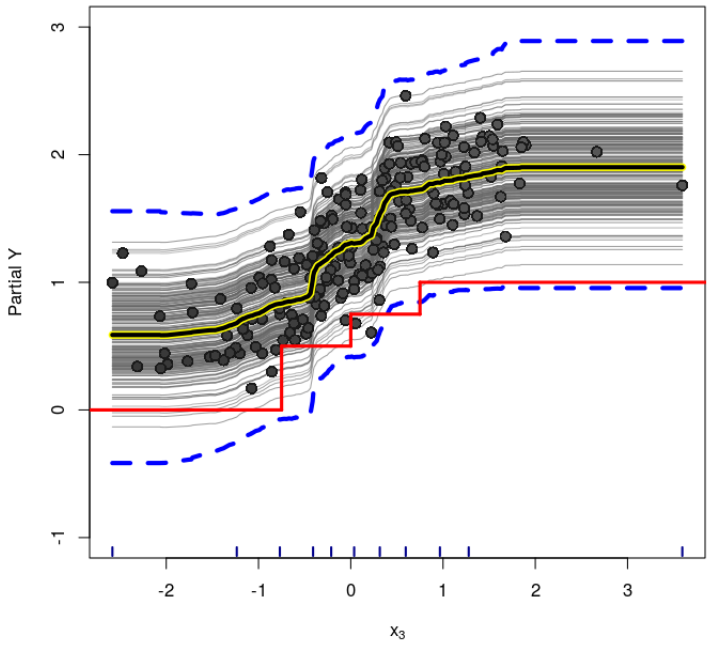

Figure B.97: *Rand-BCF model (Toy Example) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

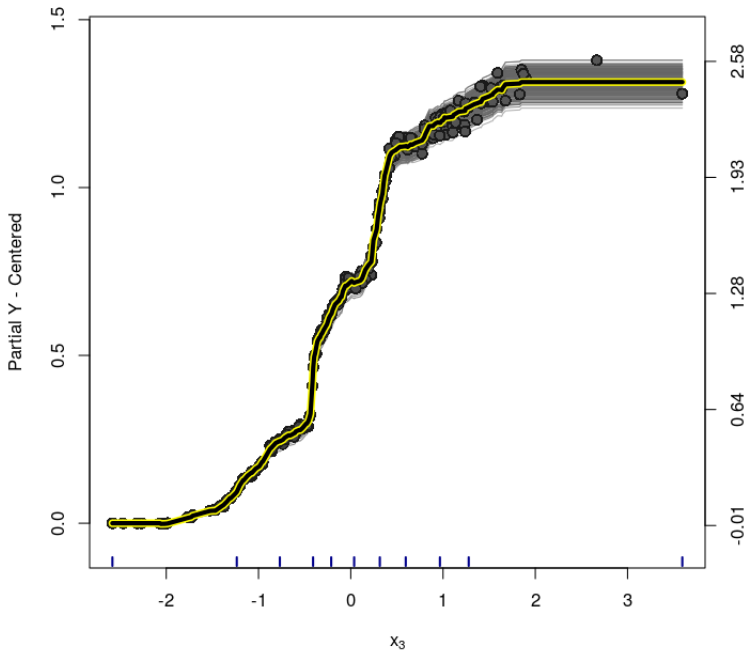

Figure B.98: *Rand-BCF model (Toy Example) - Variable x_3 - centered-ICE Plot for the treatment effect.*

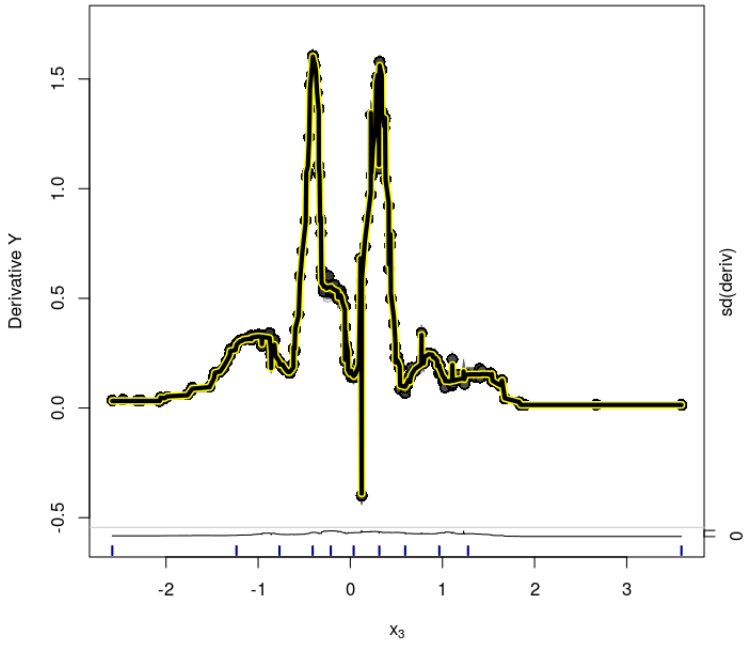

Figure B.99: *Rand-BCF model (Toy Example) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

Appendix C

Graphics - Simulation Based on Real Data

C.1 Convergence Analysis

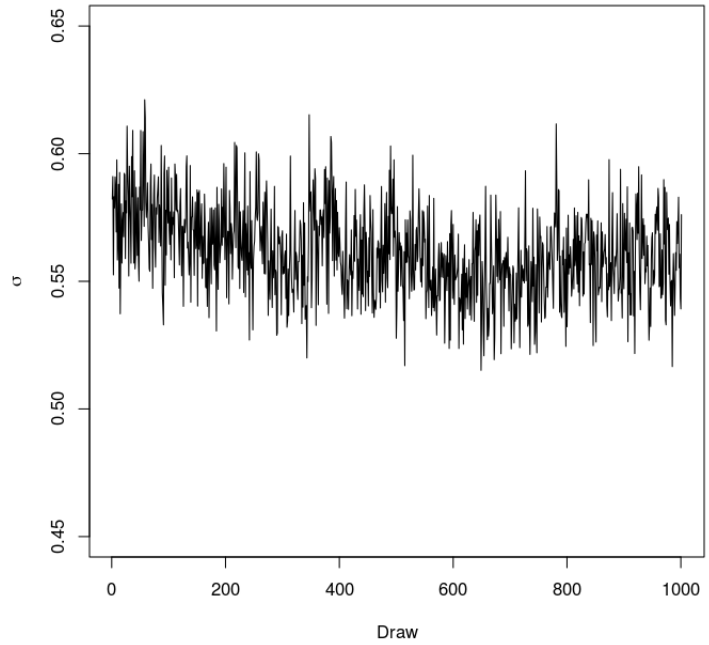

Figure C.1: *Vanilla model (Simulation Based on Real Data) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

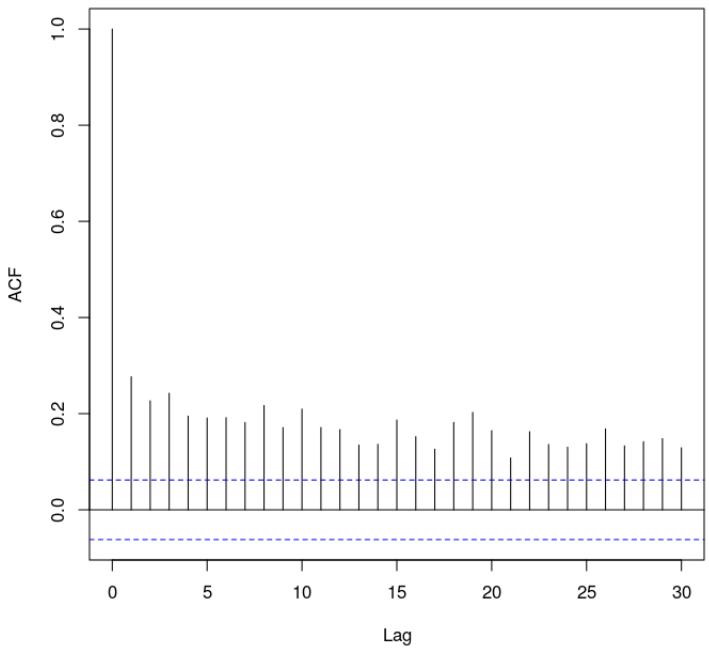

Figure C.2: *Vanilla model (Simulation Based on Real Data) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

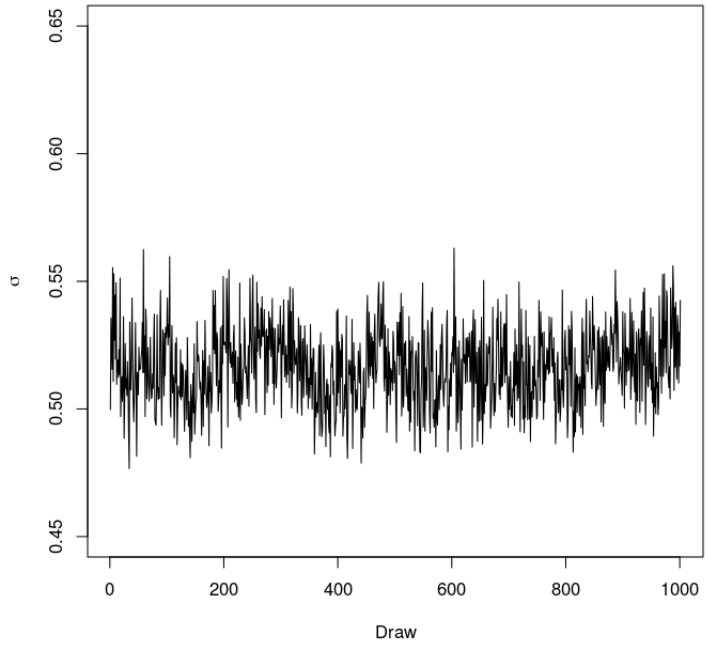

Figure C.3: *Oracle model (Simulation Based on Real Data) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

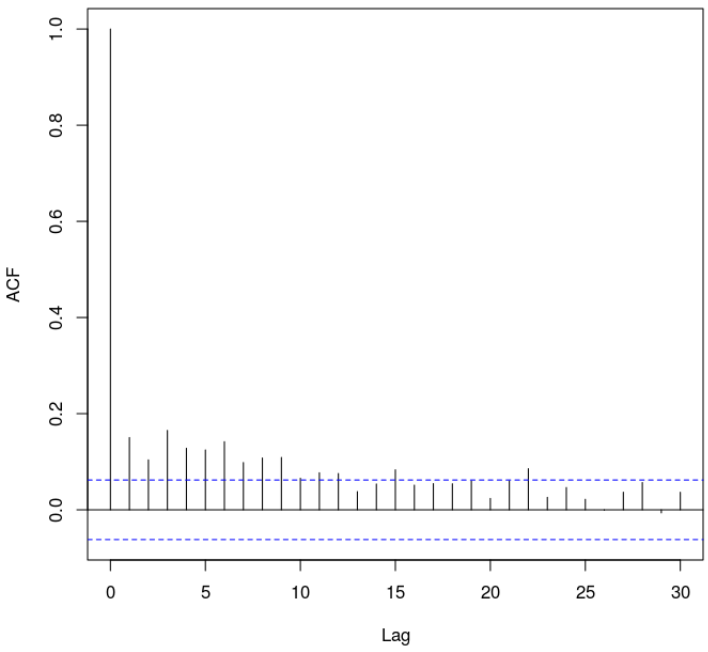

Figure C.4: *Oracle model (Simulation Based on Real Data) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

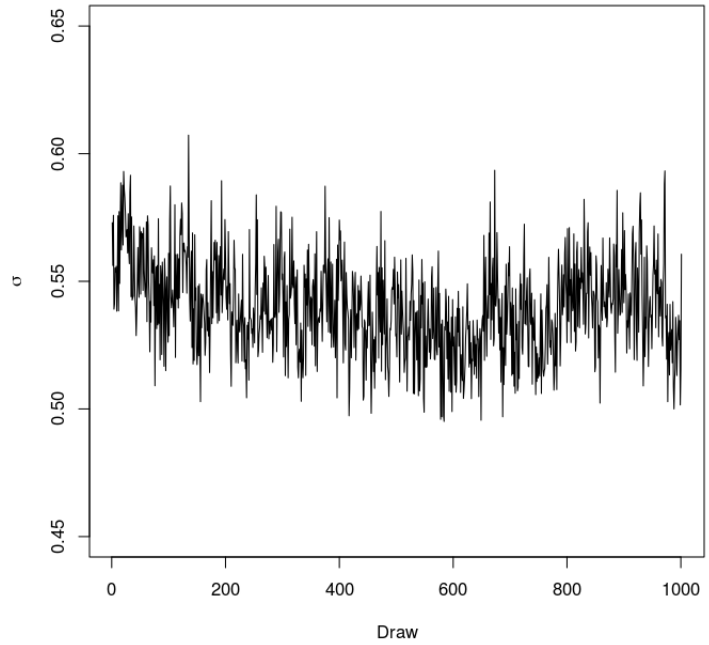

Figure C.5: *PS-BART model (Simulation Based on Real Data) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

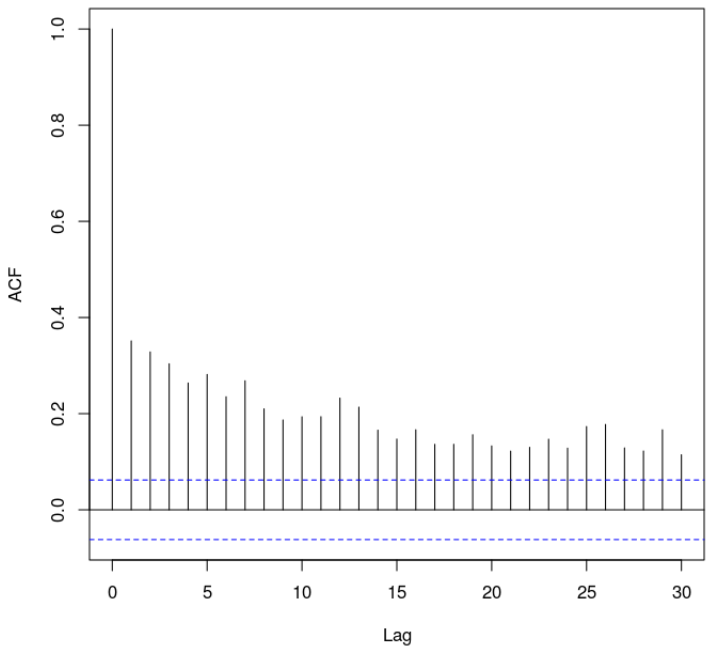

Figure C.6: *PS-BART model (Simulation Based on Real Data) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

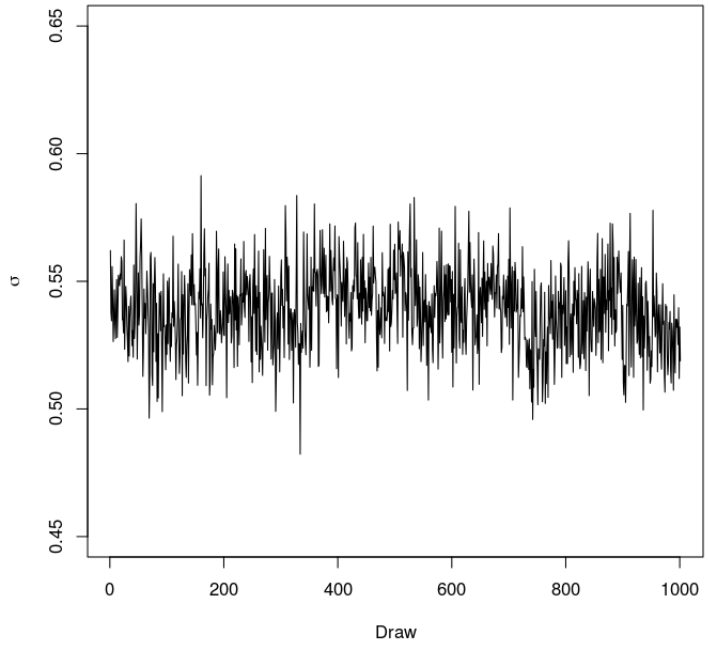

Figure C.7: *GLM-BART model (Simulation Based on Real Data) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

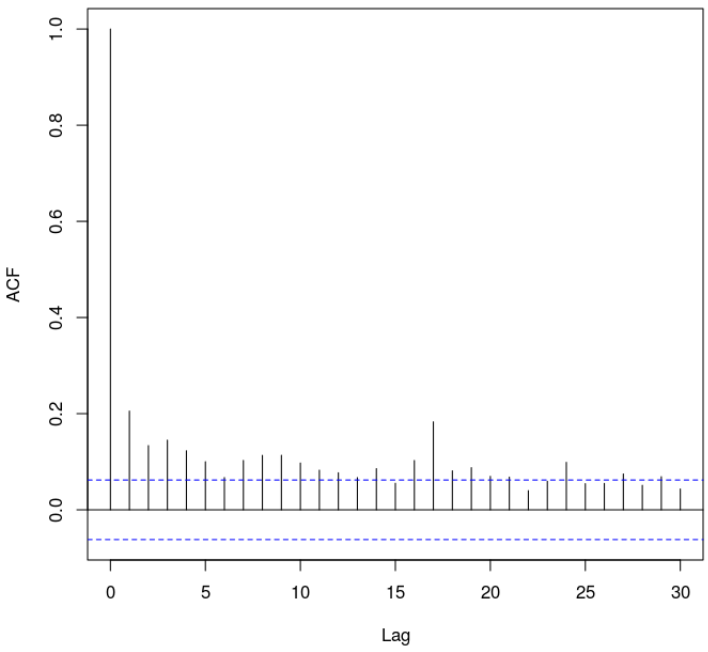

Figure C.8: *GLM-BART model (Simulation Based on Real Data) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

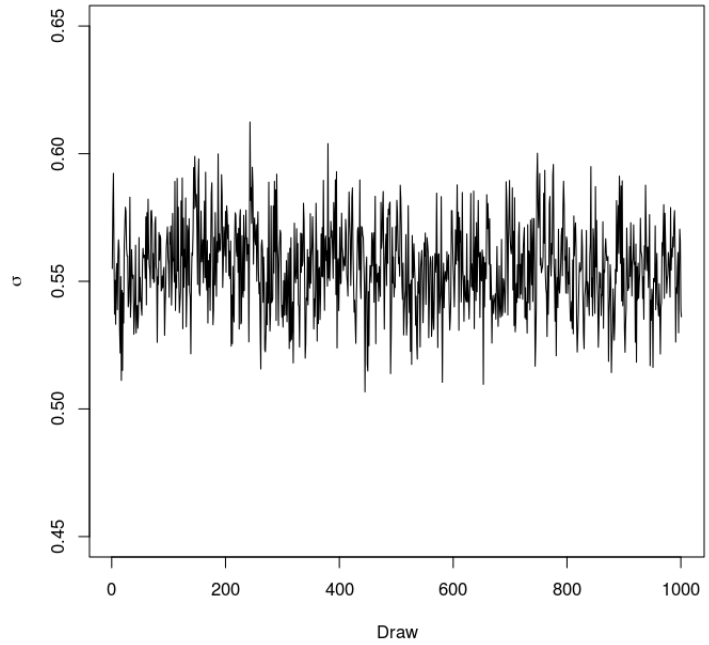

Figure C.9: *Rand-BART model (Simulation Based on Real Data) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

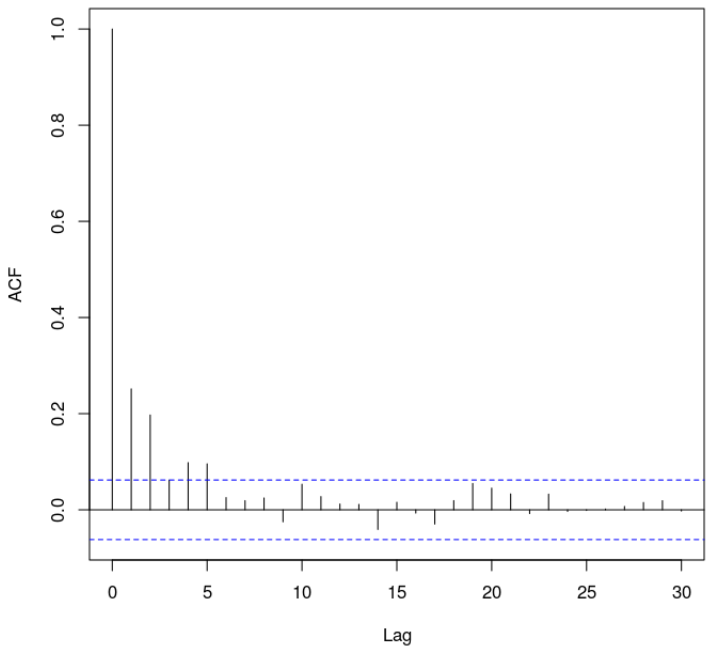

Figure C.10: *Rand-BART model (Simulation Based on Real Data) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

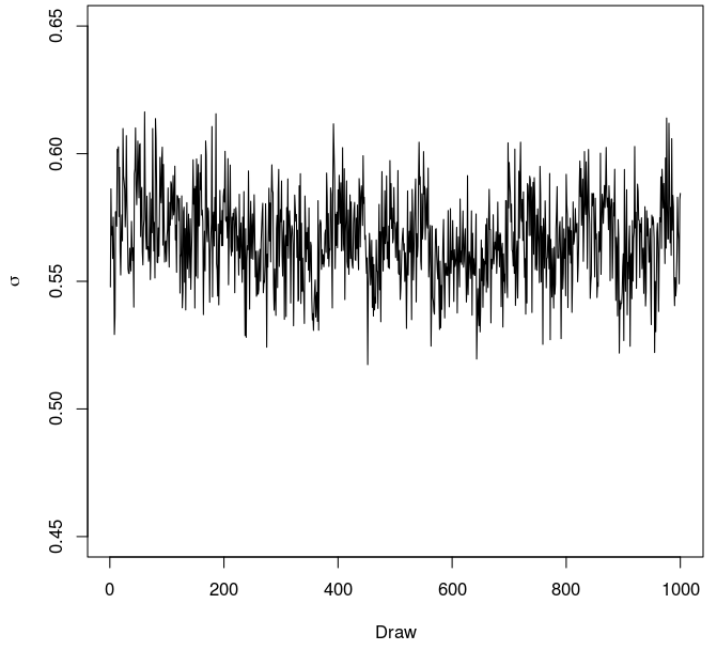

Figure C.11: *BART-BCF model (Simulation Based on Real Data) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

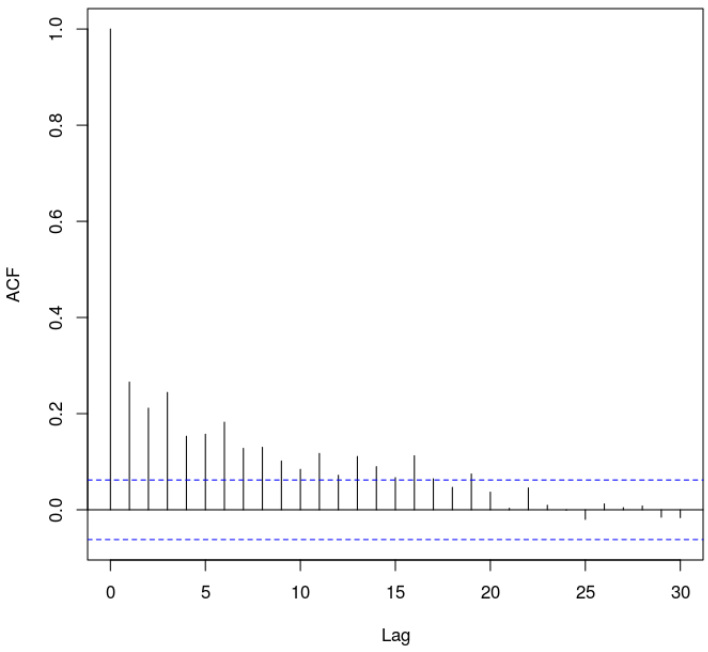

Figure C.12: *BART-BCF model (Simulation Based on Real Data) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

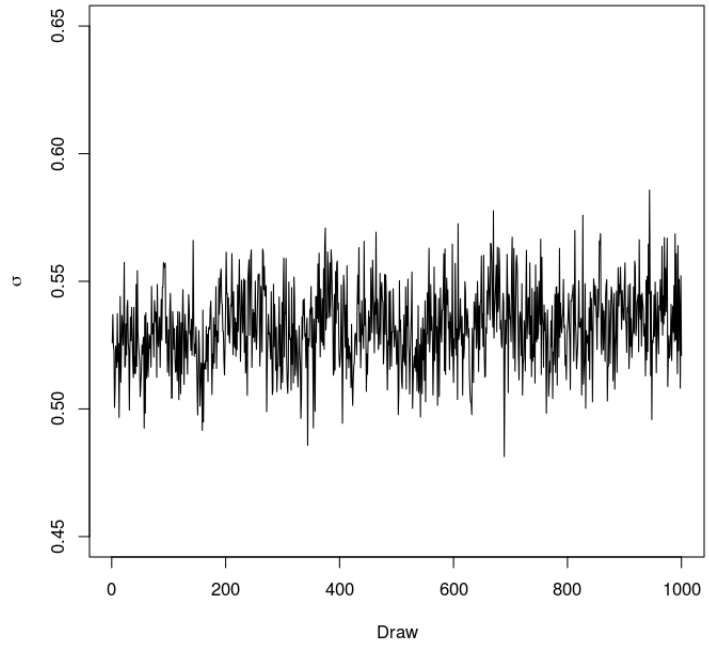

Figure C.13: *Oracle-BCF model (Simulation Based on Real Data) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

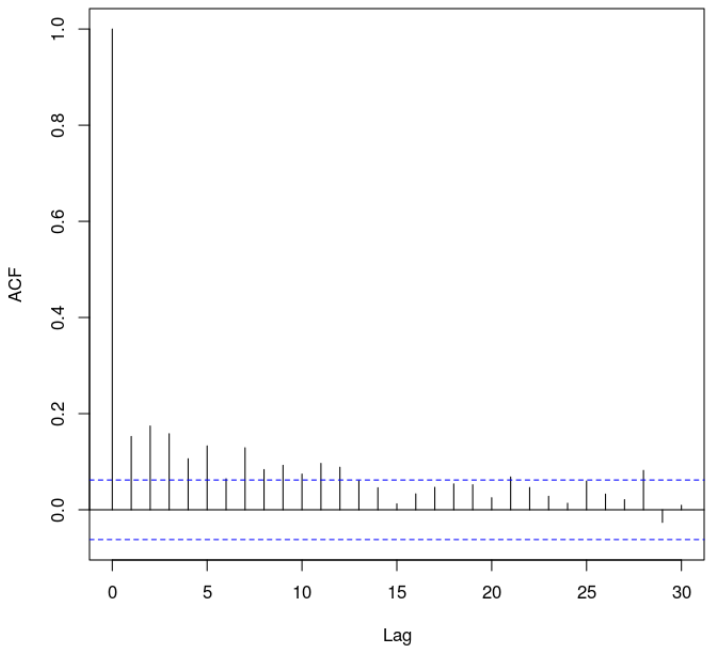

Figure C.14: *Oracle-BCF model (Simulation Based on Real Data) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

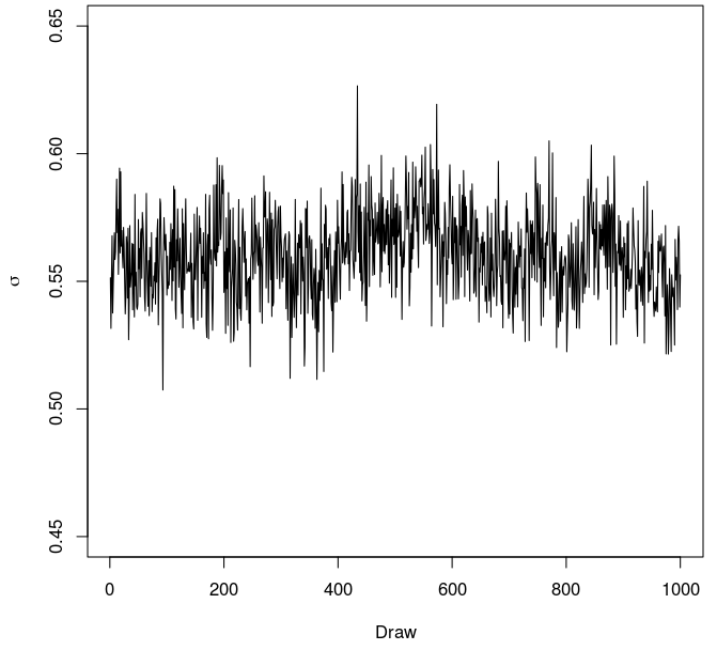

Figure C.15: *GLM-BCF model (Simulation Based on Real Data) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

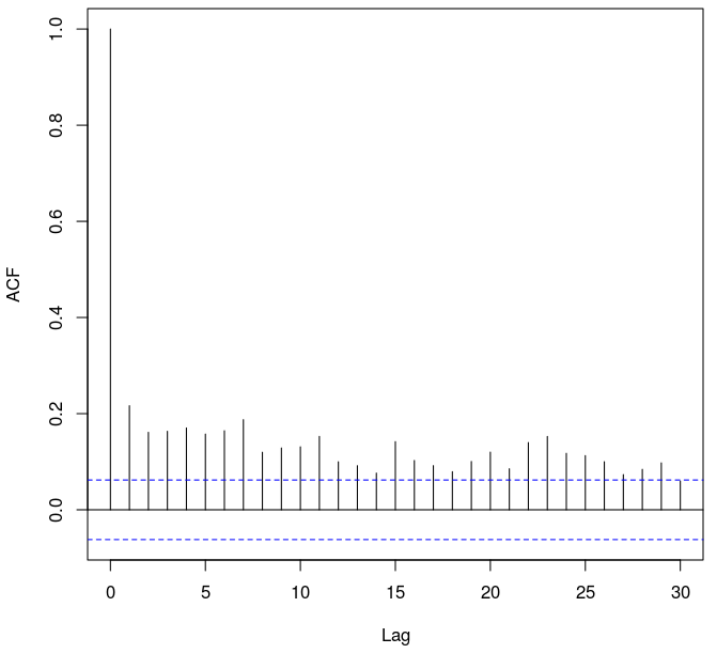

Figure C.16: *GLM-BCF model (Simulation Based on Real Data) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

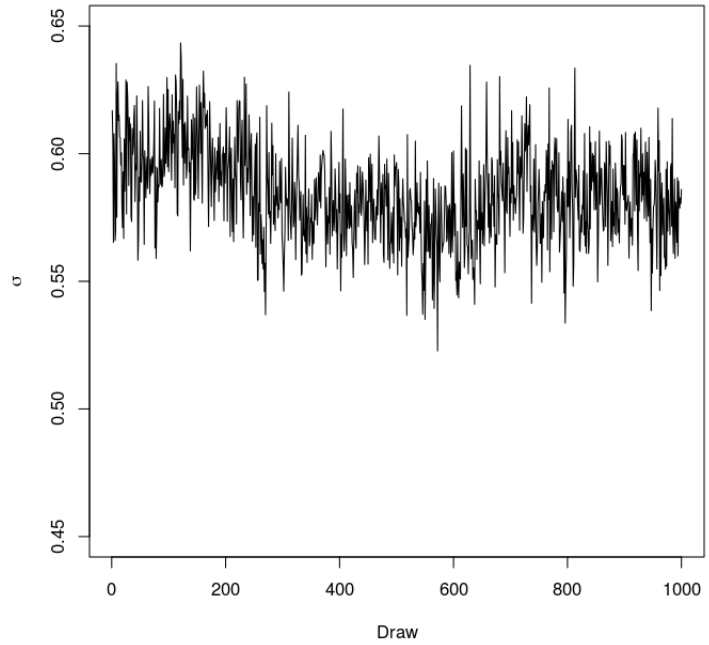

Figure C.17: *Rand-BCF model (Simulation Based on Real Data) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

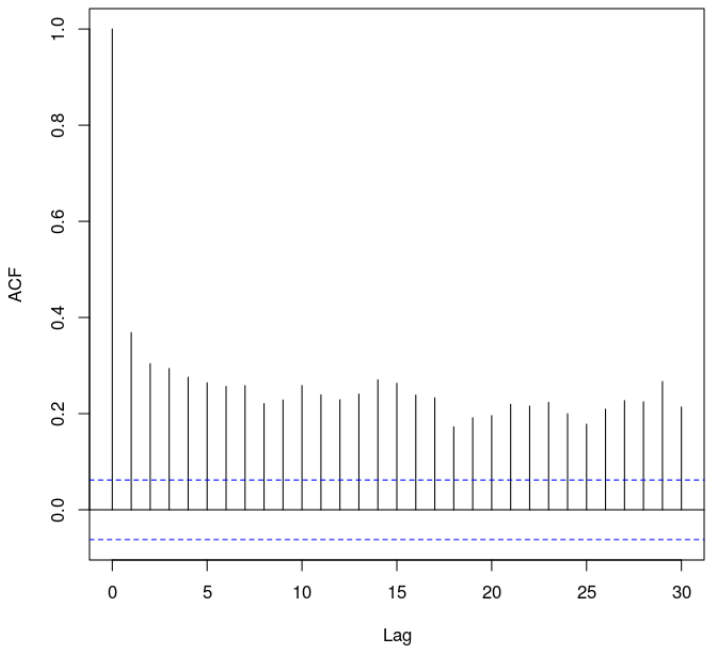

Figure C.18: *Rand-BCF model (Simulation Based on Real Data) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

C.2 ICE Plots

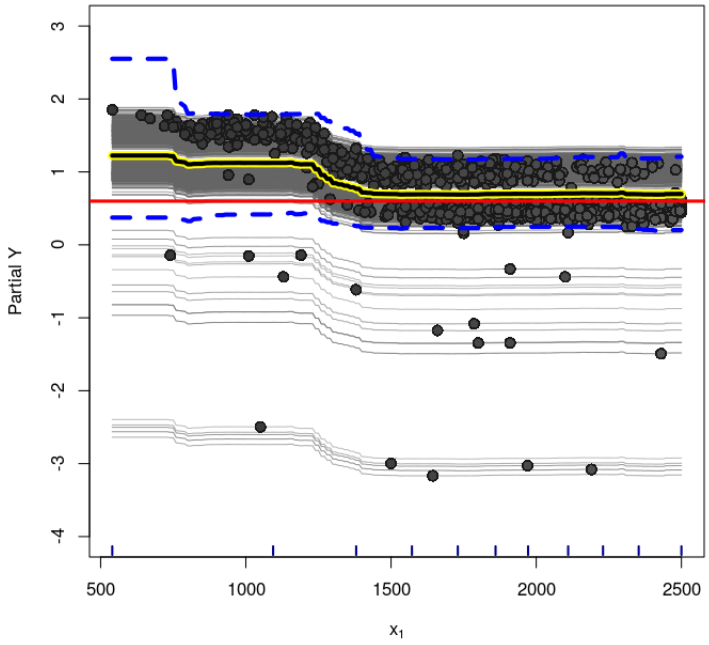

Figure C.19: *Vanilla model (Simulation Based on Real Data) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

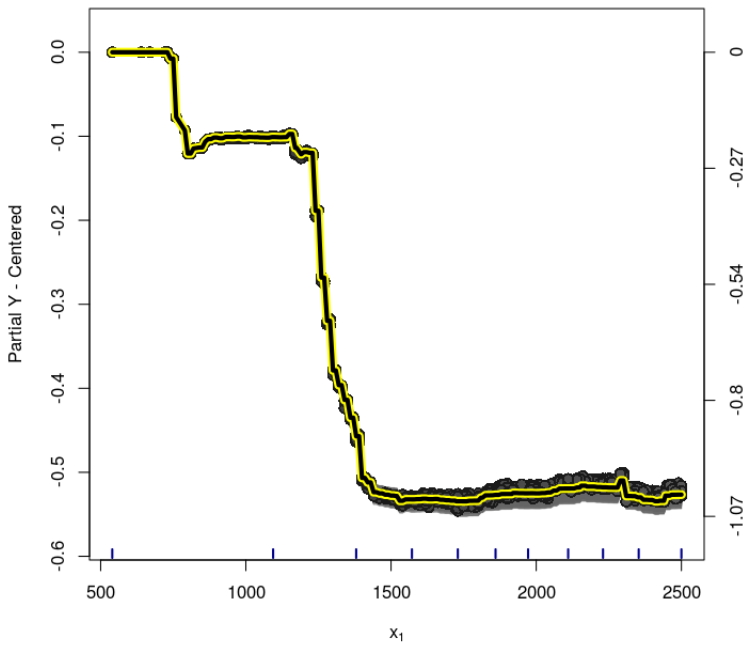

Figure C.20: *Vanilla model (Simulation Based on Real Data) - Variable x_1 - centered-ICE Plot for the treatment effect.*

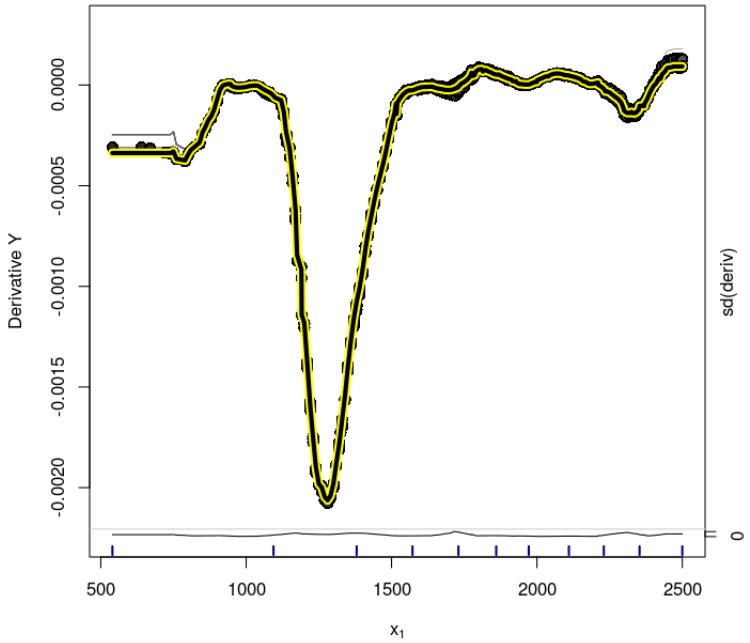

Figure C.21: *Vanilla model (Simulation Based on Real Data) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

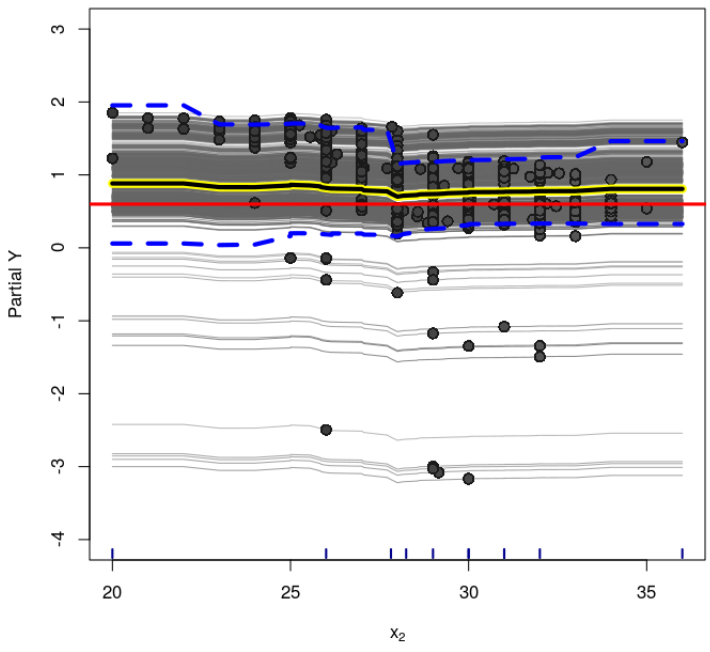

Figure C.22: *Vanilla model (Simulation Based on Real Data) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

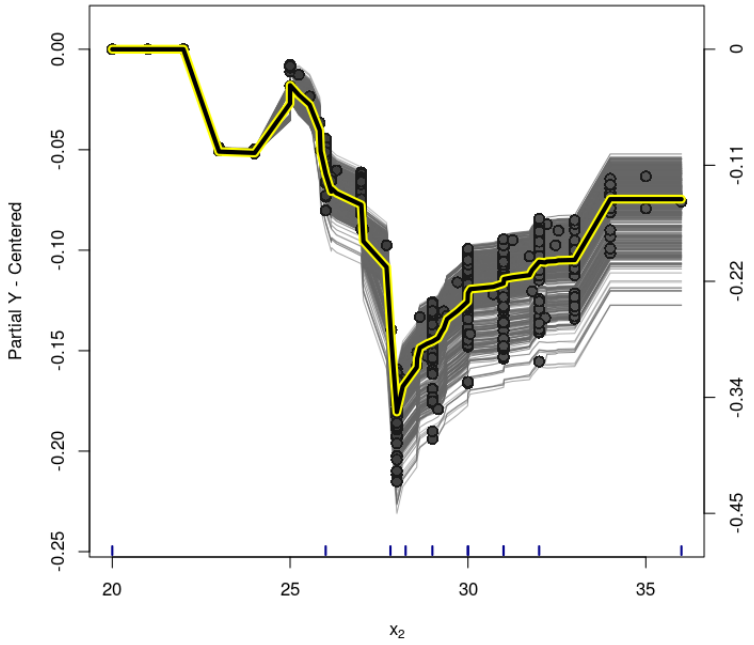

Figure C.23: *Vanilla model (Simulation Based on Real Data) - Variable x_2 - centered-ICE Plot for the treatment effect.*

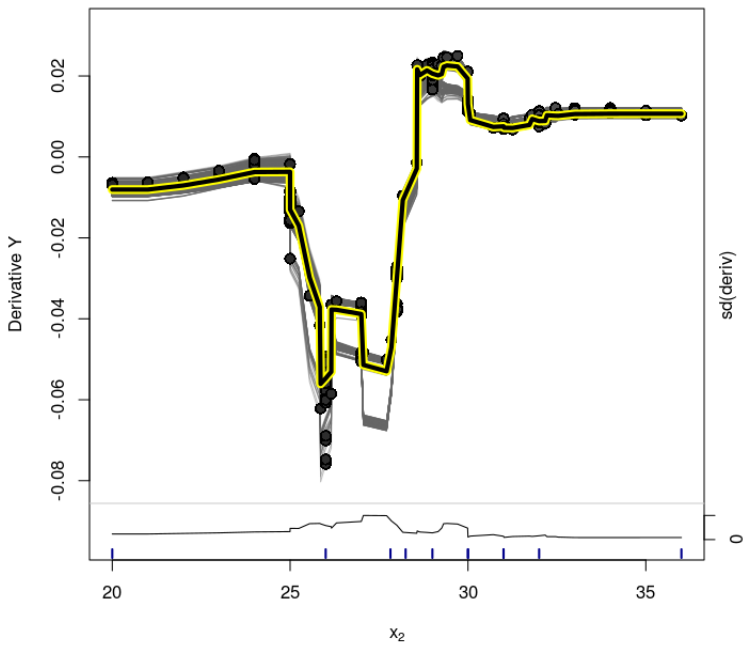

Figure C.24: *Vanilla model (Simulation Based on Real Data) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

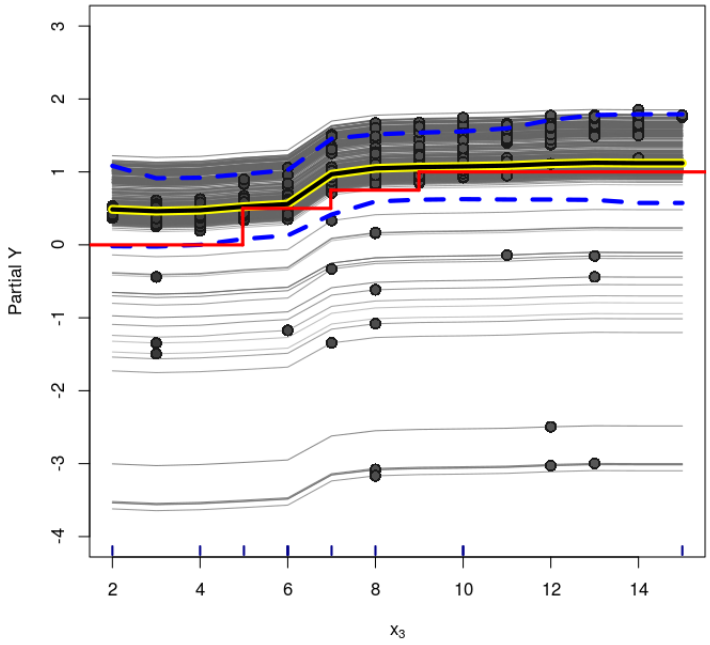

Figure C.25: *Vanilla model (Simulation Based on Real Data) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

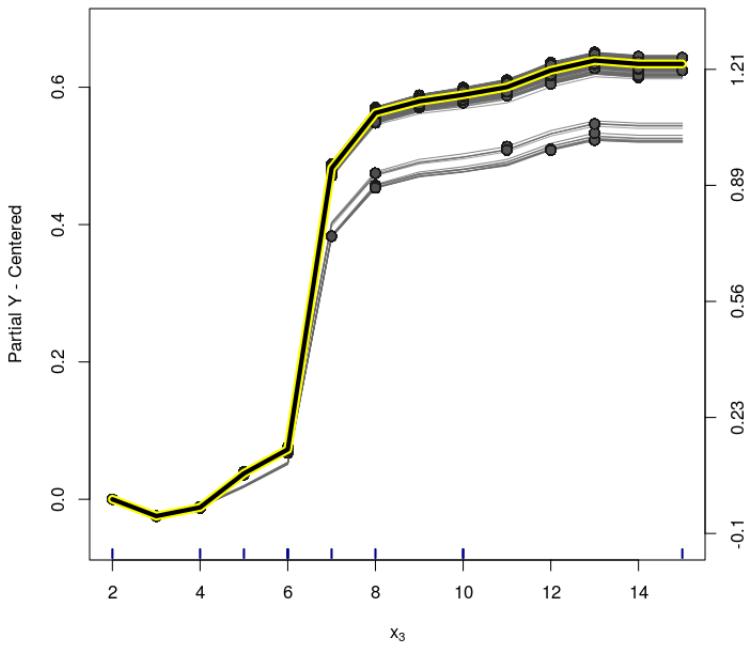

Figure C.26: *Vanilla model (Simulation Based on Real Data) - Variable x_3 - centered-ICE Plot for the treatment effect.*

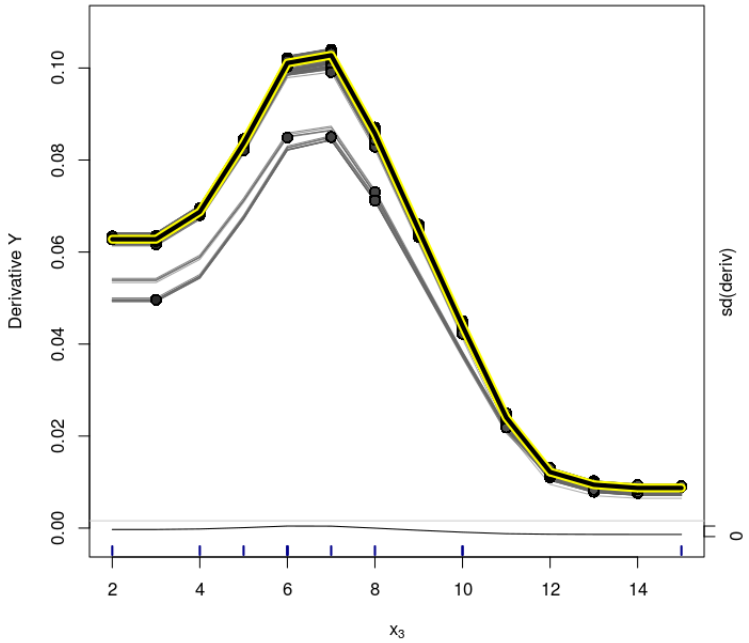

Figure C.27: *Vanilla model (Simulation Based on Real Data) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

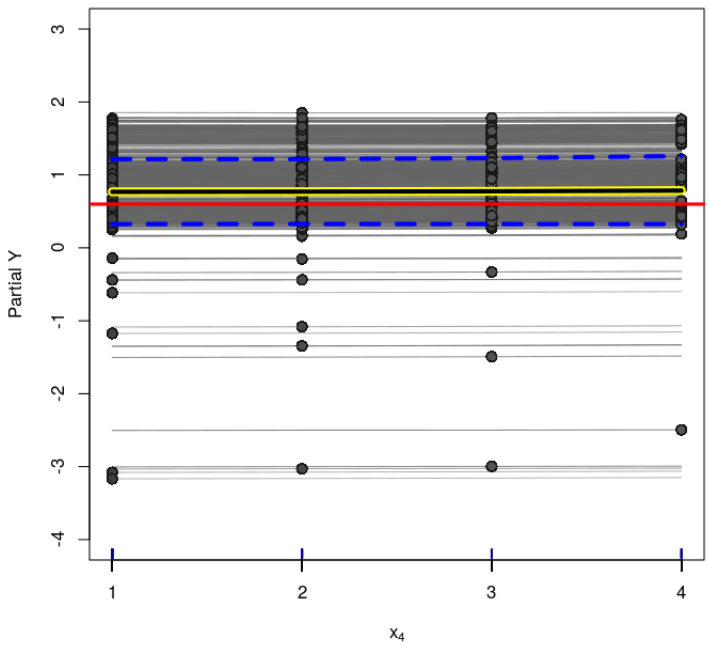

Figure C.28: *Vanilla model (Simulation Based on Real Data) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

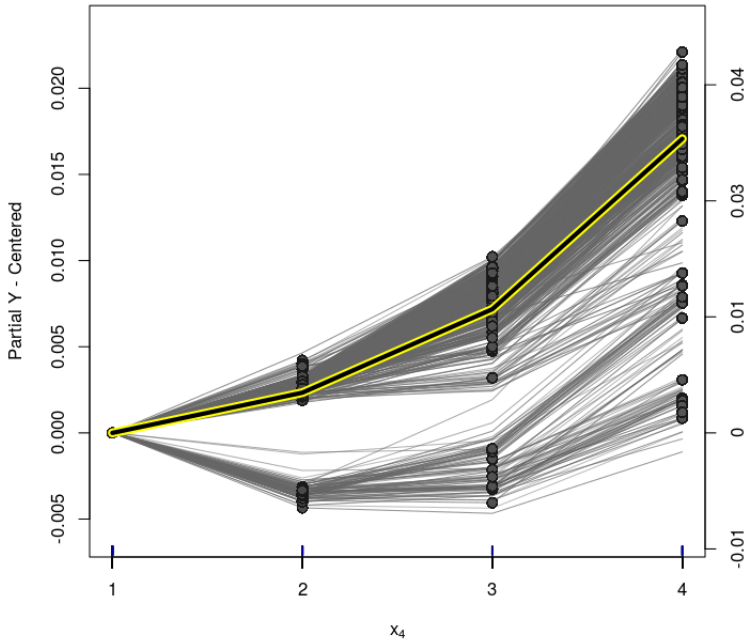

Figure C.29: *Vanilla model (Simulation Based on Real Data) - Variable x_4 - centered-ICE Plot for the treatment effect.*

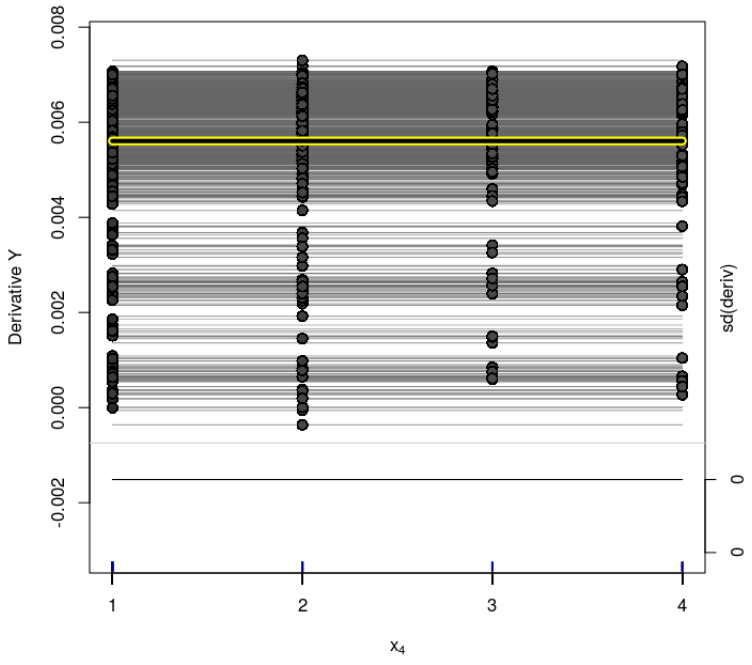

Figure C.30: *Vanilla model (Simulation Based on Real Data) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

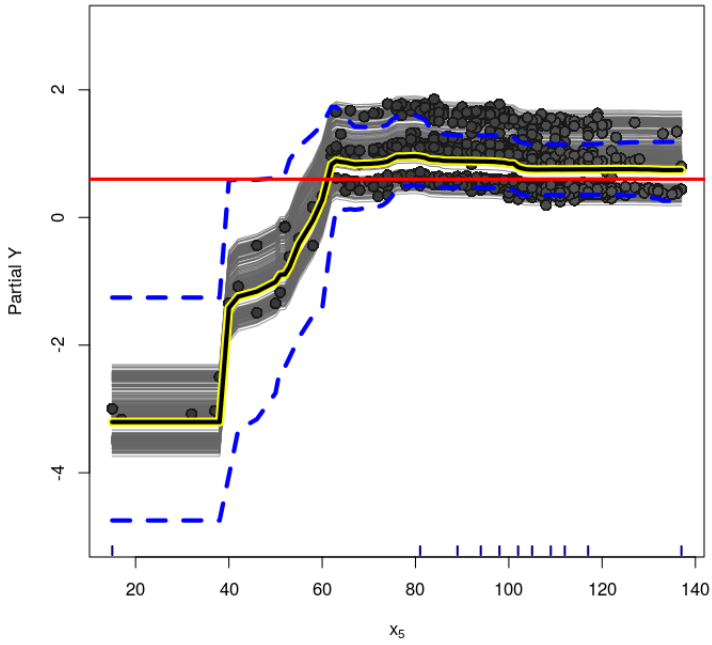

Figure C.31: *Vanilla model (Simulation Based on Real Data) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

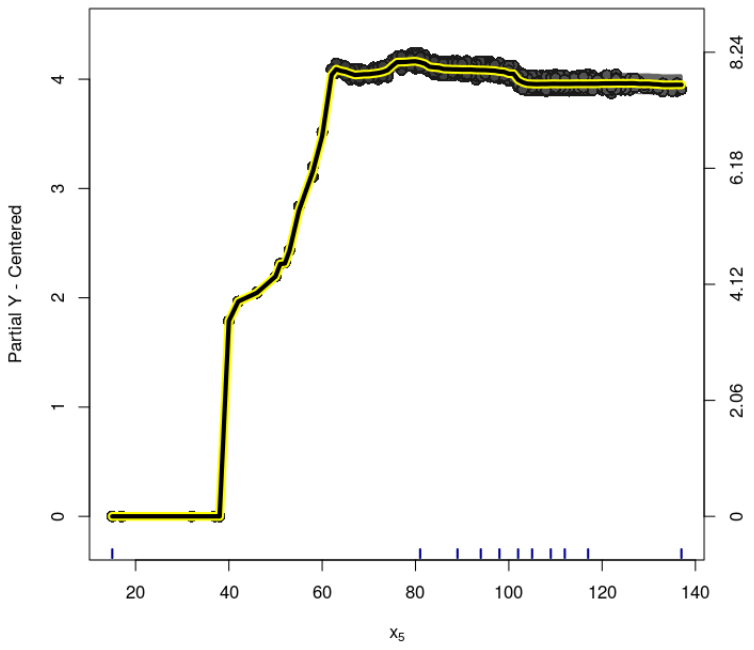

Figure C.32: *Vanilla model (Simulation Based on Real Data) - Variable x_5 - centered-ICE Plot for the treatment effect.*

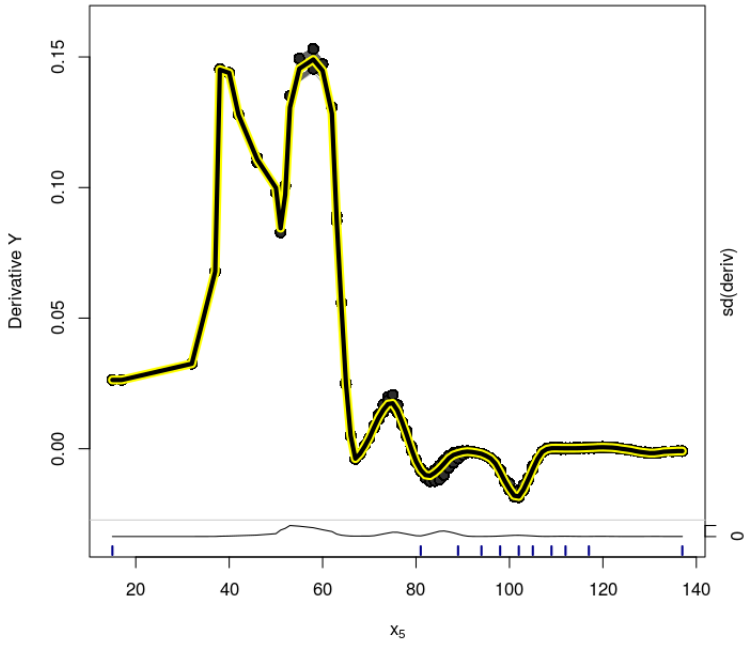

Figure C.33: *Vanilla model (Simulation Based on Real Data) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

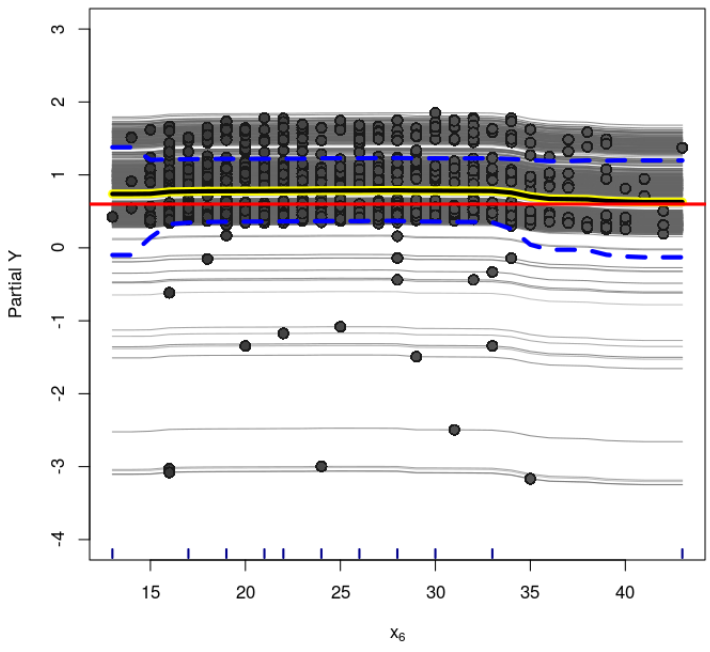

Figure C.34: *Vanilla model (Simulation Based on Real Data) - Variable x_6 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

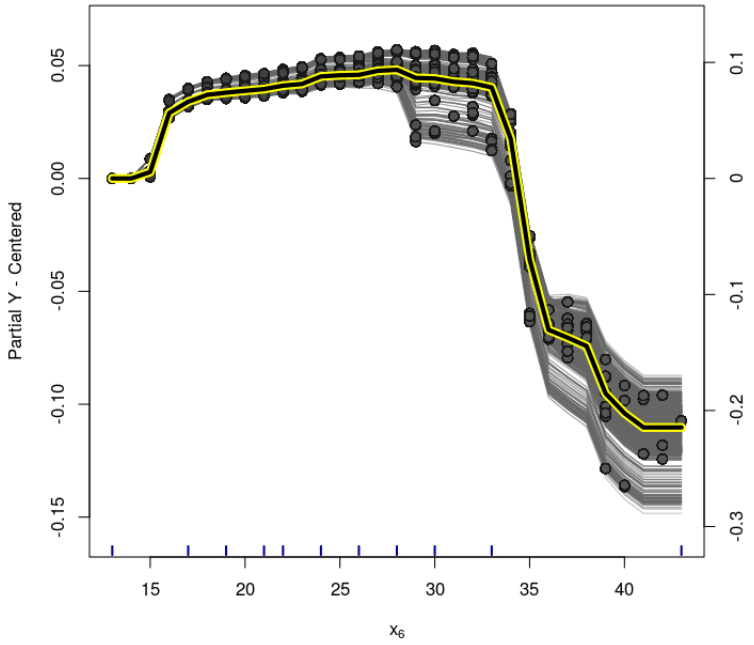

Figure C.35: *Vanilla model (Simulation Based on Real Data) - Variable x_6 - centered-ICE Plot for the treatment effect.*

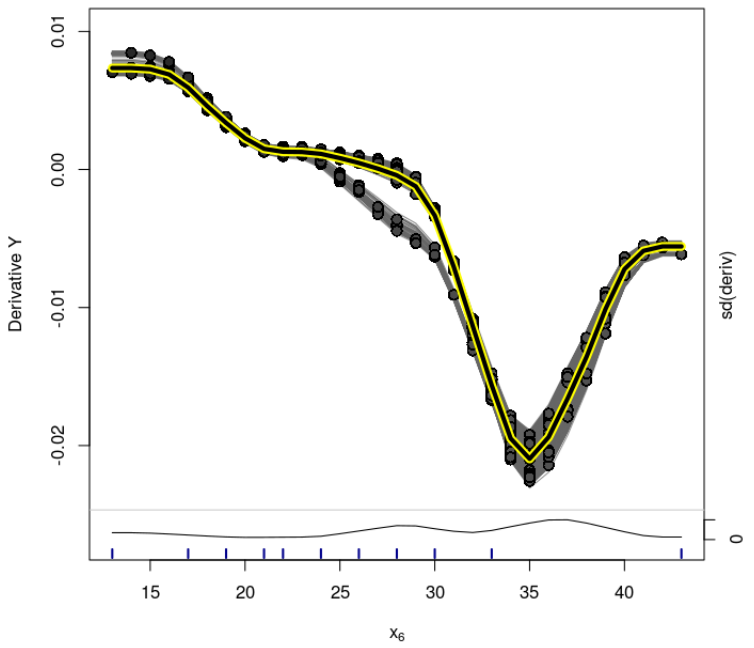

Figure C.36: *Vanilla model (Simulation Based on Real Data) - Variable x_6 - d-ICE Plot for the treatment effect estimates.*

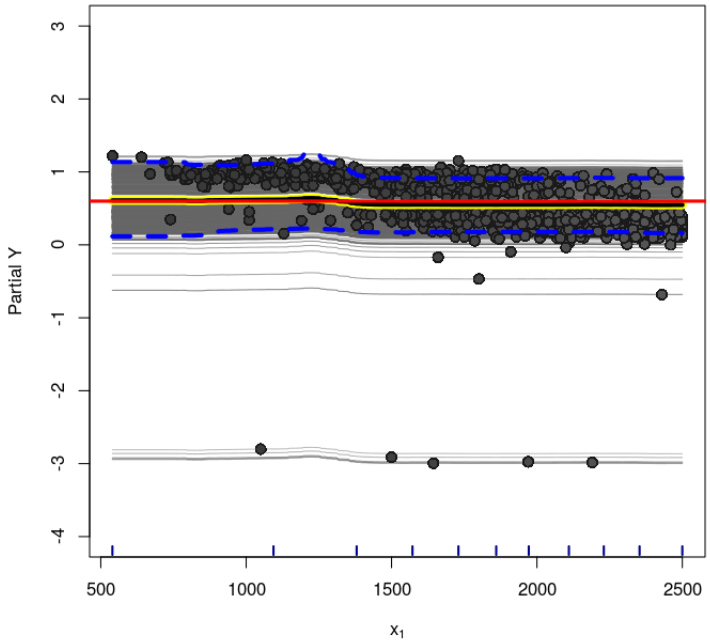

Figure C.37: *Oracle model (Simulation Based on Real Data) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

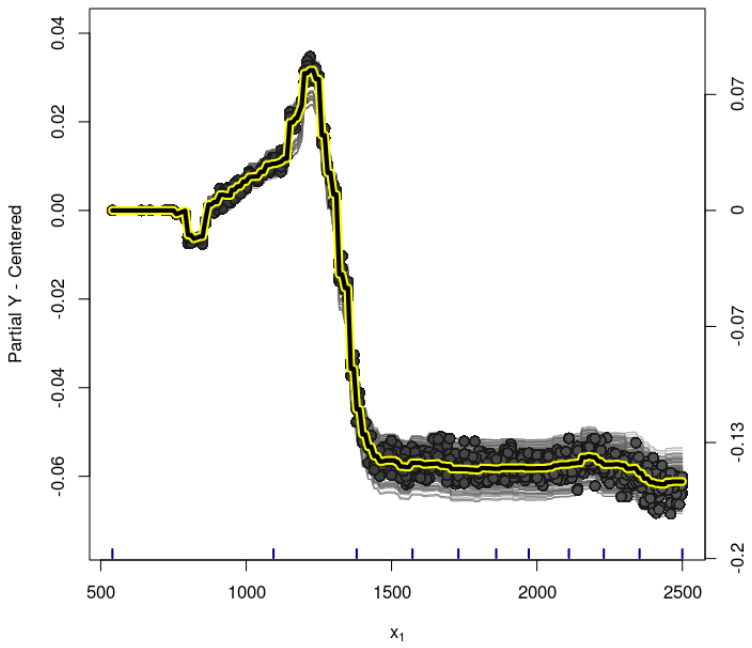

Figure C.38: *Oracle model (Simulation Based on Real Data) - Variable x_1 - centered-ICE Plot for the treatment effect.*

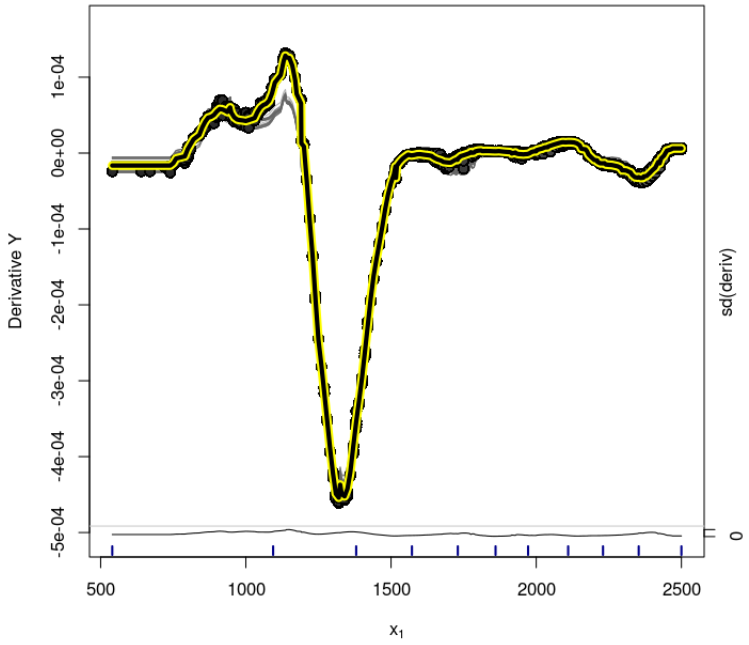

Figure C.39: *Oracle model (Simulation Based on Real Data) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

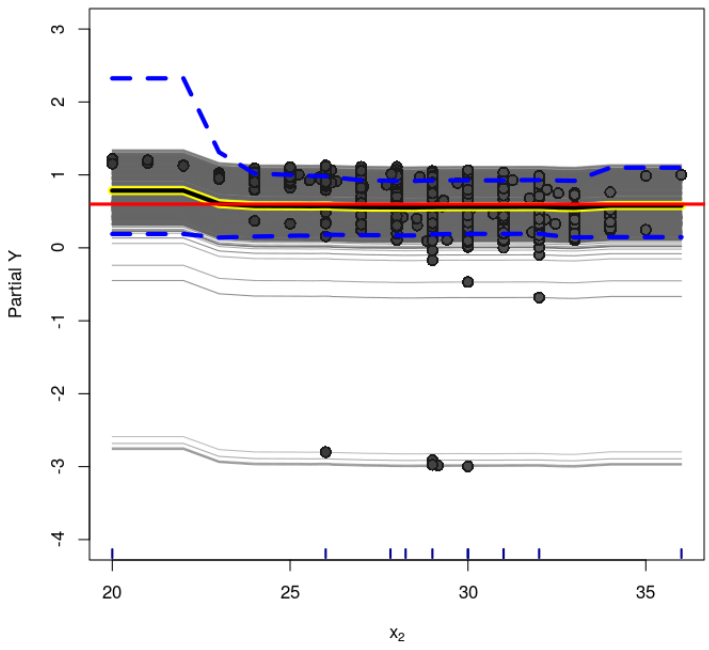

Figure C.40: *Oracle model (Simulation Based on Real Data) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

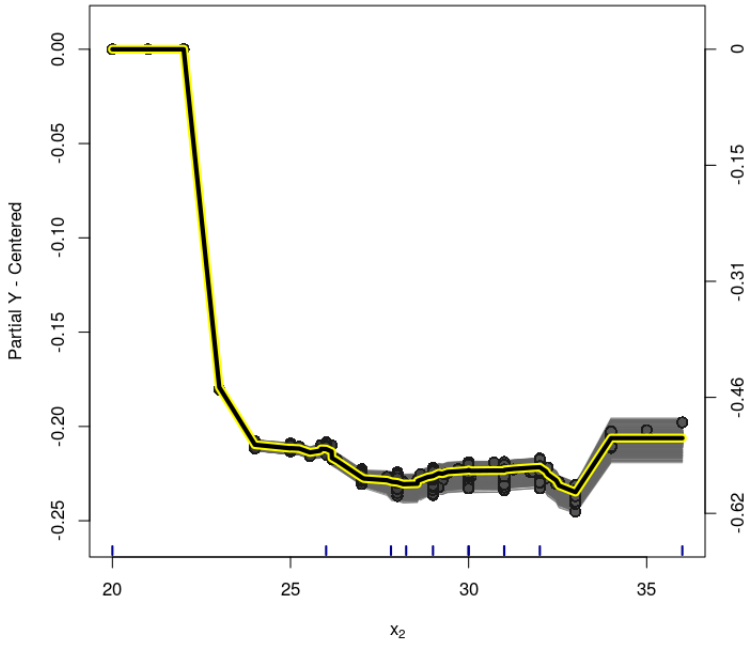

Figure C.41: *Oracle model (Simulation Based on Real Data) - Variable x_2 - centered-ICE Plot for the treatment effect.*

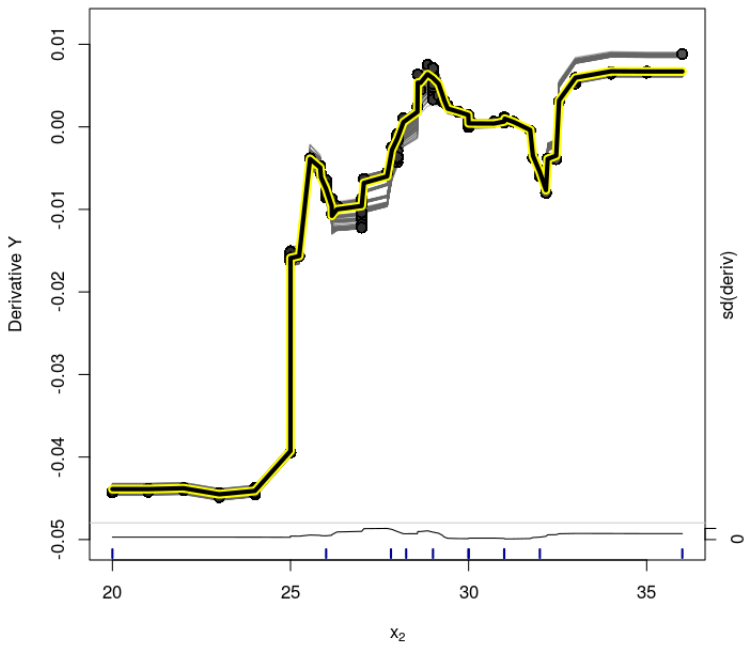

Figure C.42: *Oracle model (Simulation Based on Real Data) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

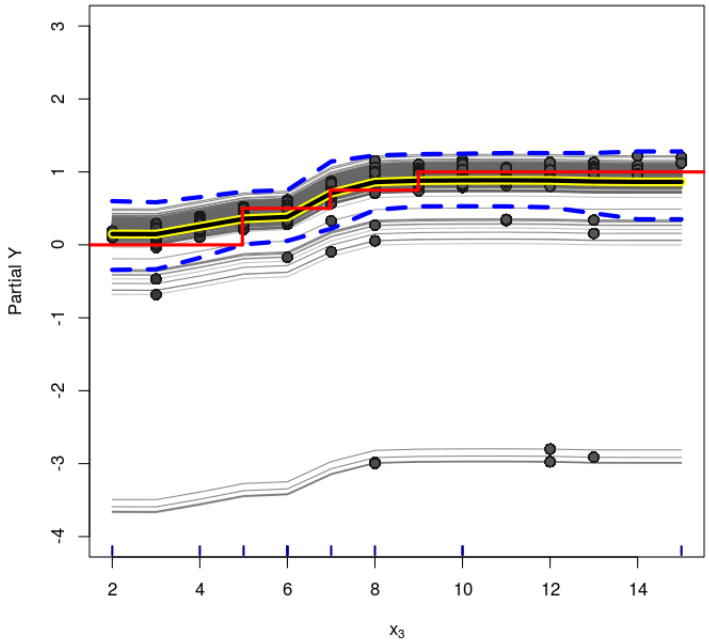

Figure C.43: *Oracle model (Simulation Based on Real Data) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

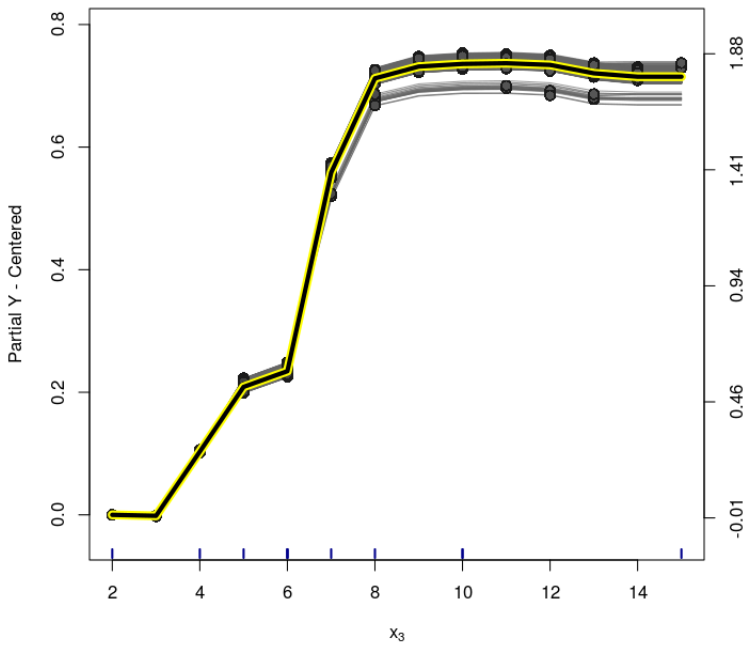

Figure C.44: *Oracle model (Simulation Based on Real Data) - Variable x_3 - centered-ICE Plot for the treatment effect.*

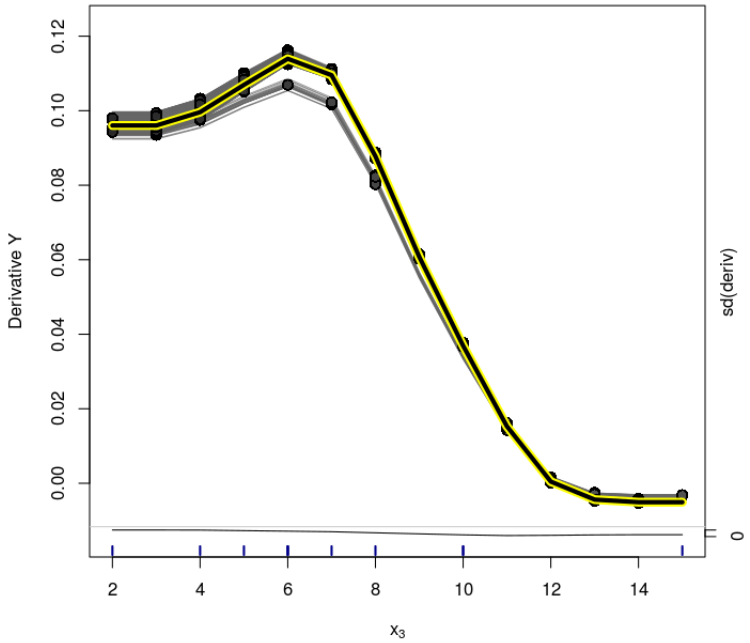

Figure C.45: *Oracle model (Simulation Based on Real Data) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

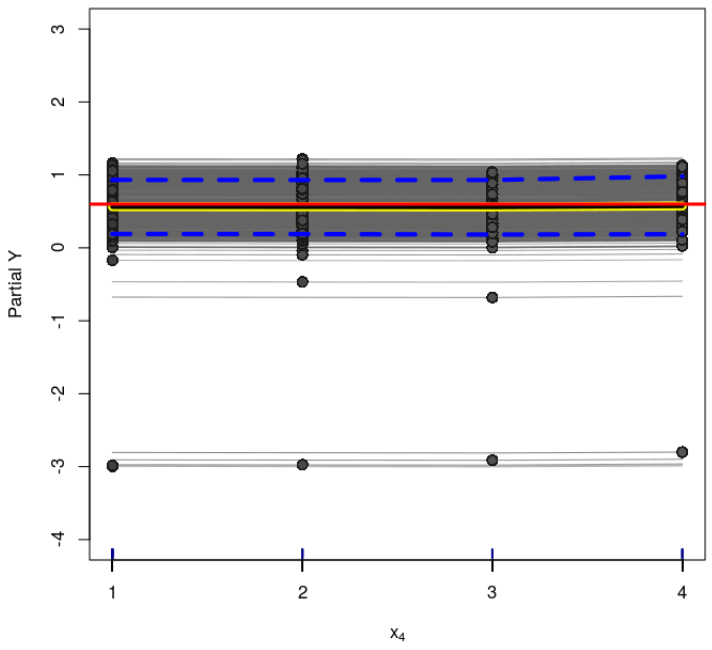

Figure C.46: *Oracle model (Simulation Based on Real Data) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

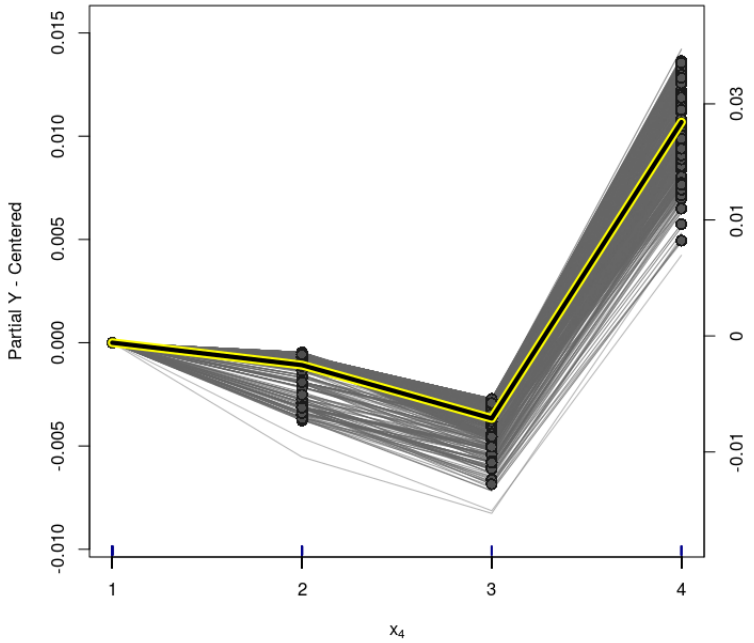

Figure C.47: *Oracle model (Simulation Based on Real Data) - Variable x_4 - centered-ICE Plot for the treatment effect.*

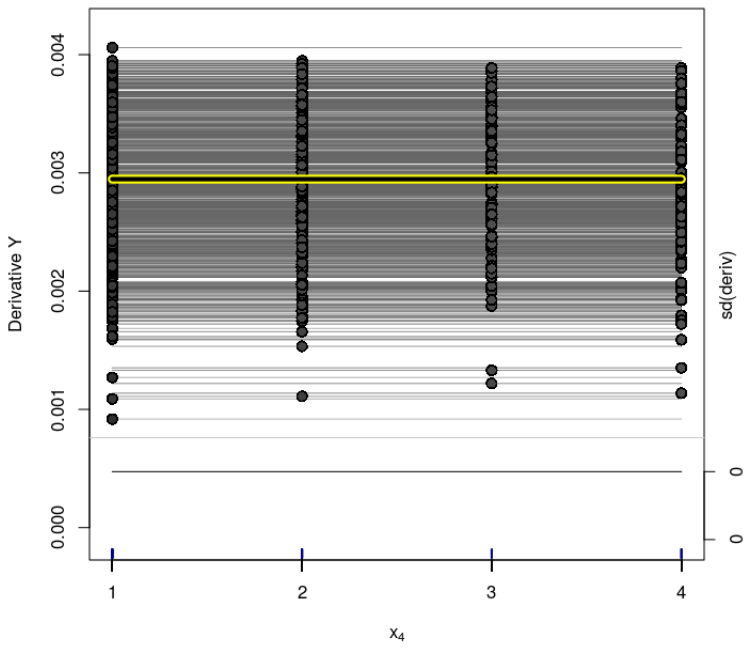

Figure C.48: *Oracle model (Simulation Based on Real Data) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

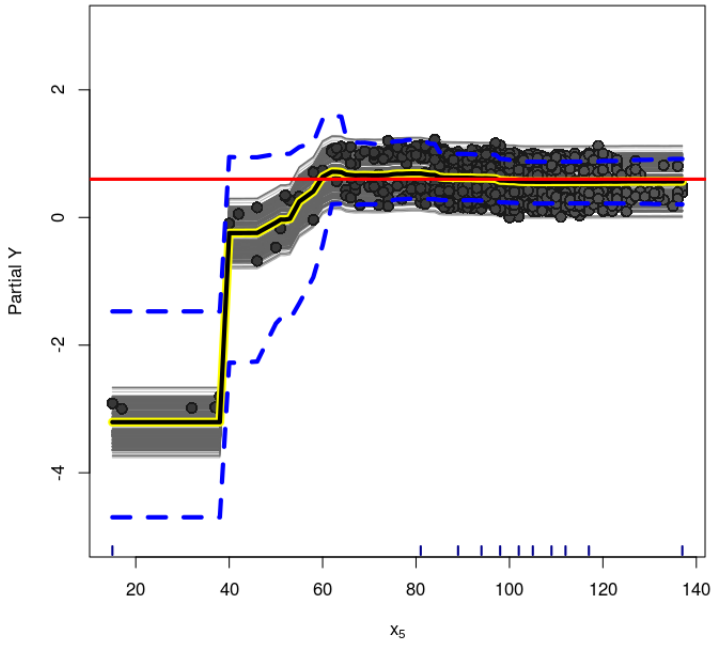

Figure C.49: *Oracle model (Simulation Based on Real Data) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

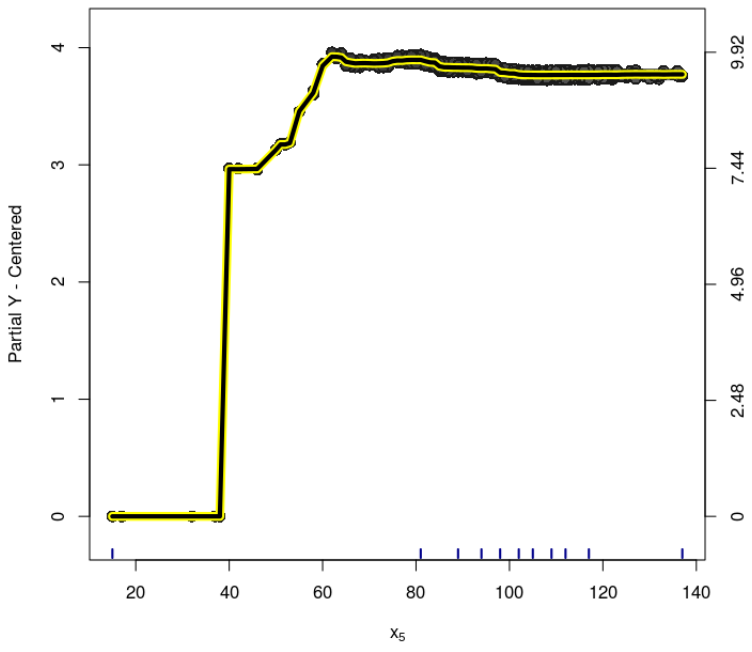

Figure C.50: *Oracle model (Simulation Based on Real Data) - Variable x_5 - centered-ICE Plot for the treatment effect.*

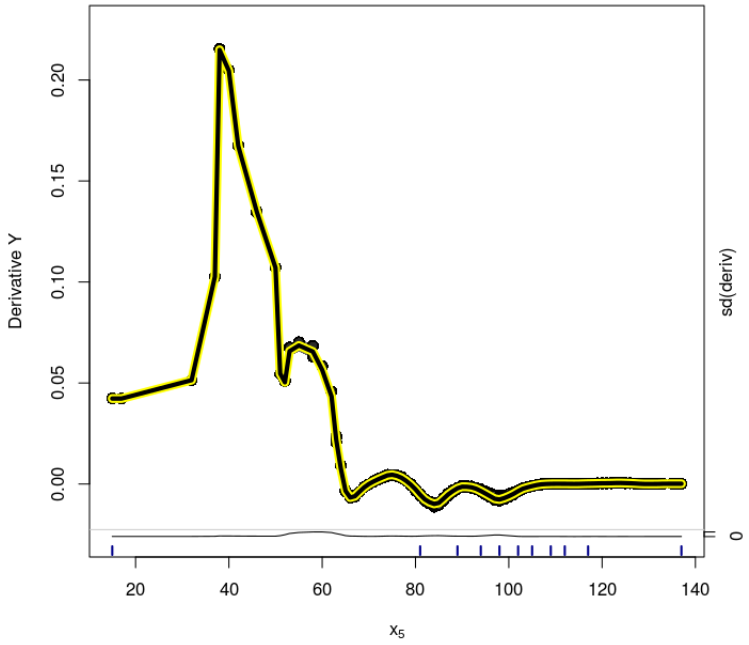

Figure C.51: *Oracle model (Simulation Based on Real Data) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

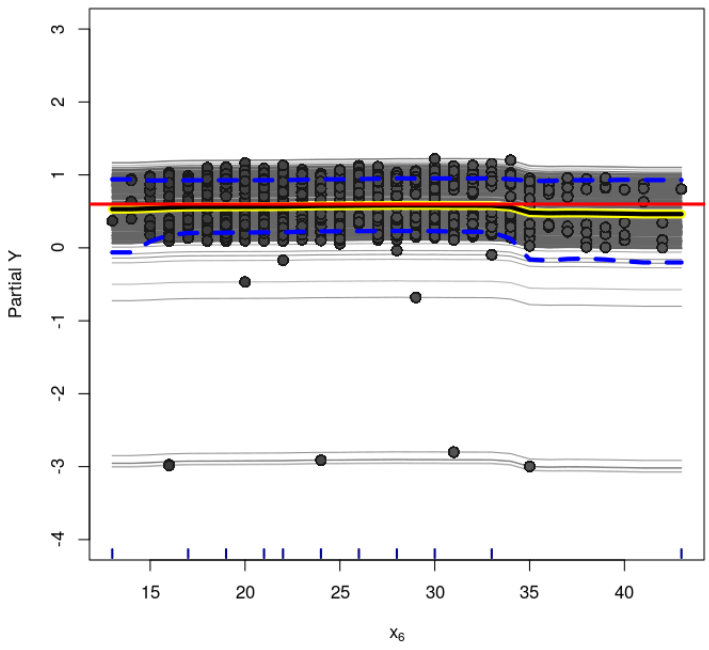

Figure C.52: *Oracle model (Simulation Based on Real Data) - Variable x_6 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

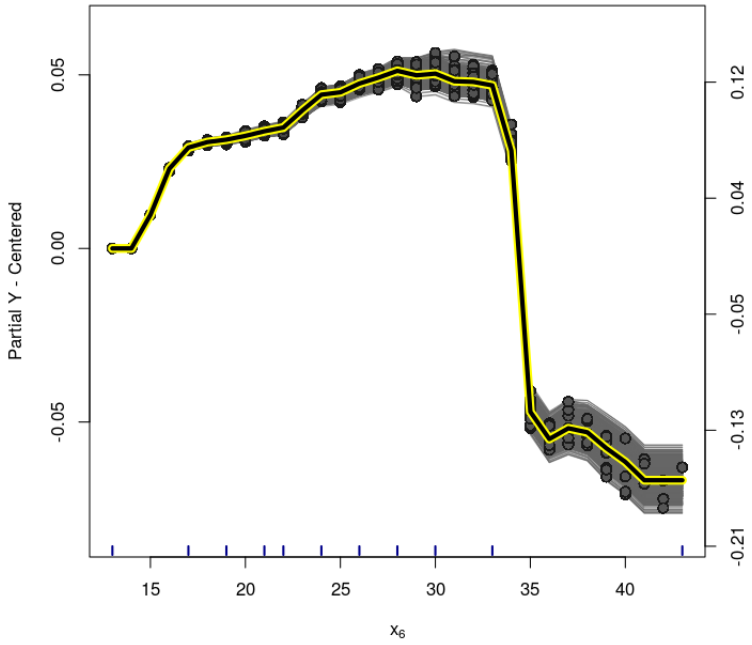

Figure C.53: *Oracle model (Simulation Based on Real Data) - Variable x_6 - centered-ICE Plot for the treatment effect.*

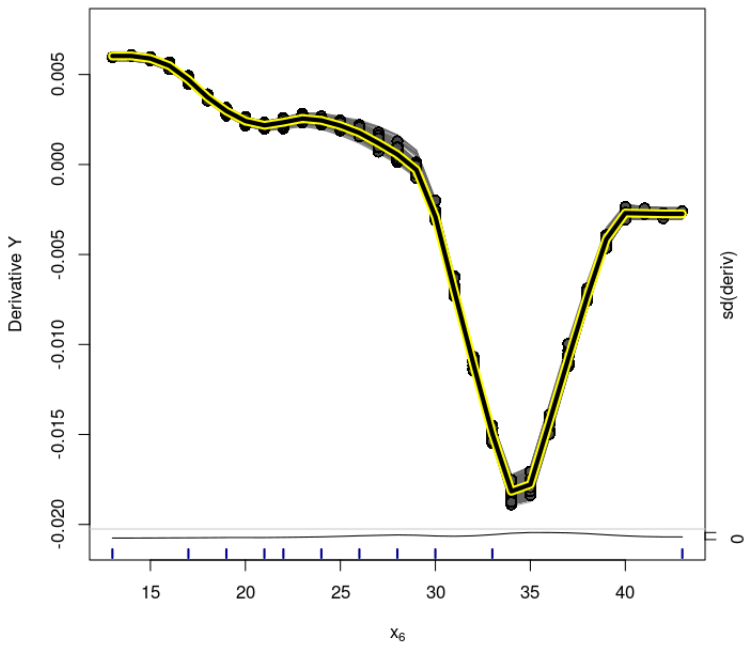

Figure C.54: *Oracle model (Simulation Based on Real Data) - Variable x_6 - d-ICE Plot for the treatment effect estimates.*

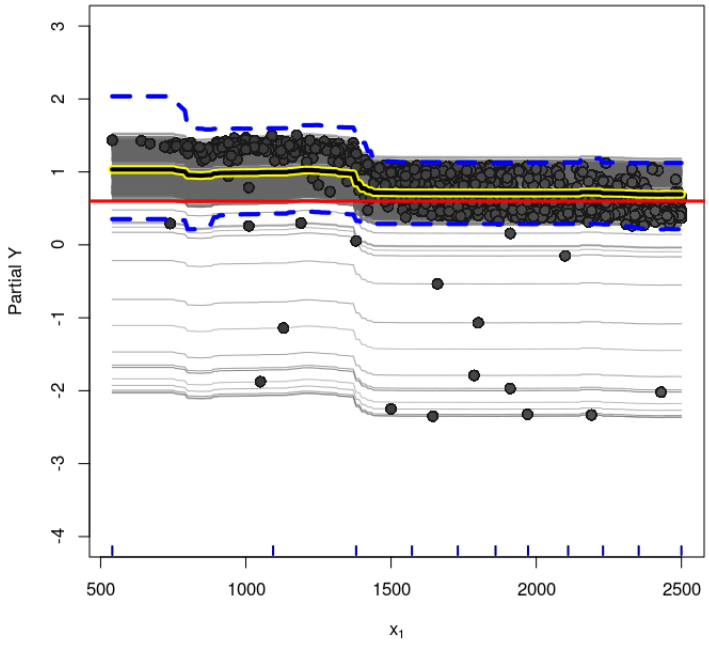

Figure C.55: *PS-BART model (Simulation Based on Real Data) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

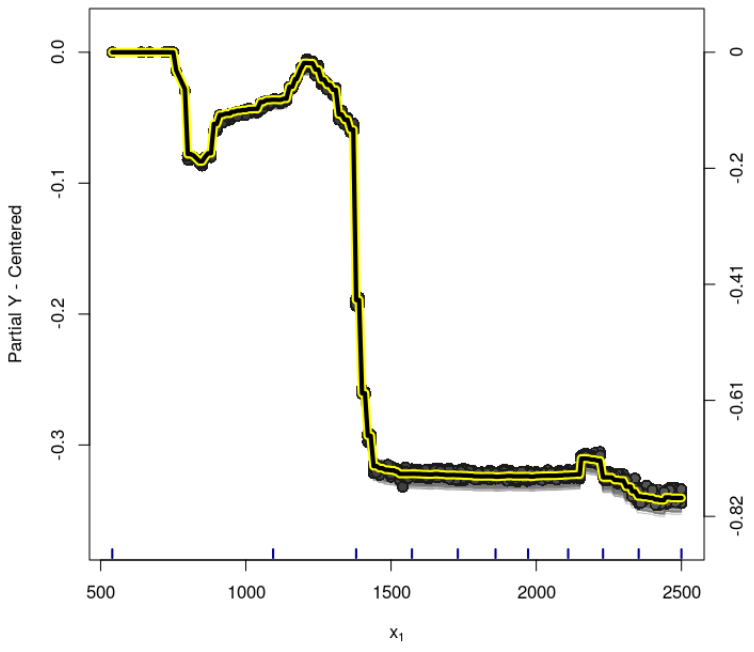

Figure C.56: *PS-BART model (Simulation Based on Real Data) - Variable x_1 - centered-ICE Plot for the treatment effect.*

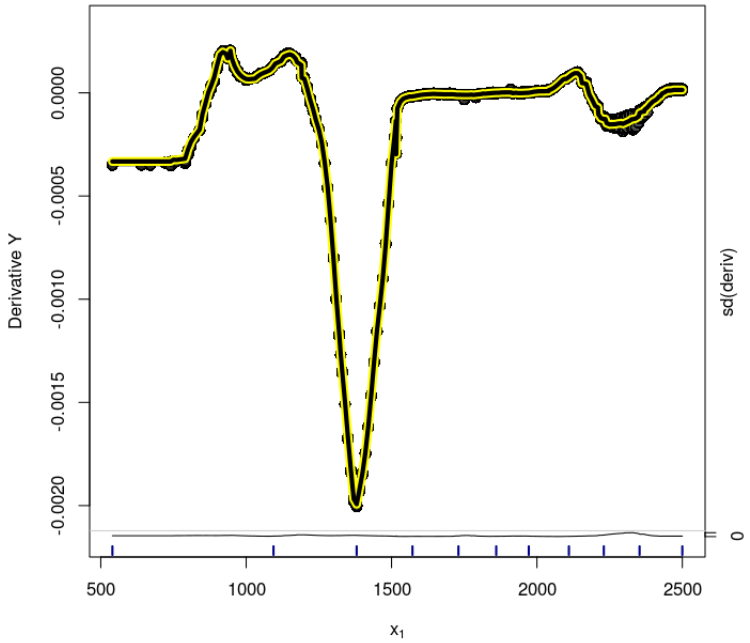

Figure C.57: *PS-BART model (Simulation Based on Real Data) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

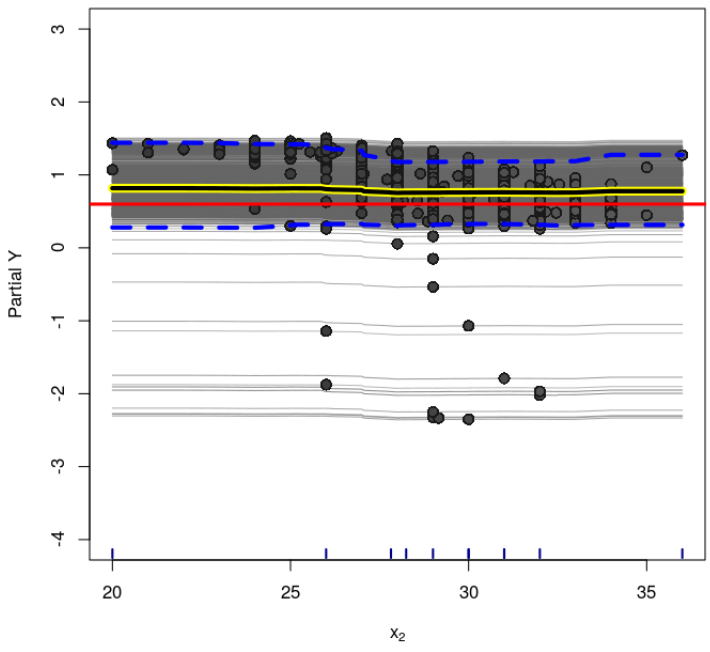

Figure C.58: *PS-BART model (Simulation Based on Real Data) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

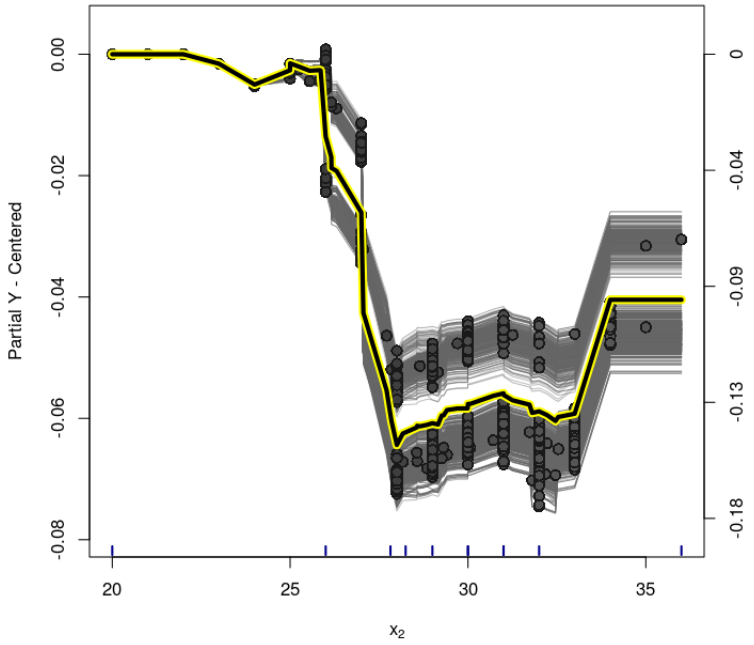

Figure C.59: *PS-BART model (Simulation Based on Real Data) - Variable x_2 - centered-ICE Plot for the treatment effect.*

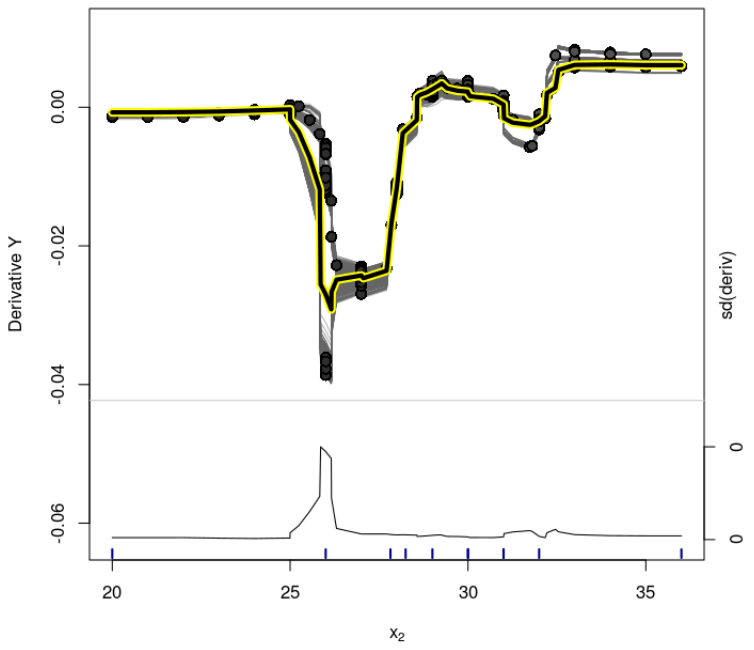

Figure C.60: *PS-BART model (Simulation Based on Real Data) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

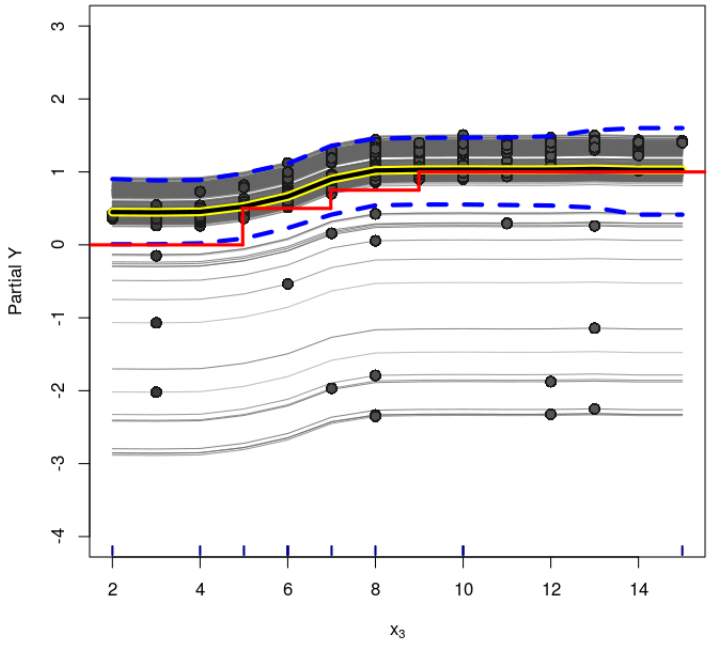

Figure C.61: *PS-BART model (Simulation Based on Real Data) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

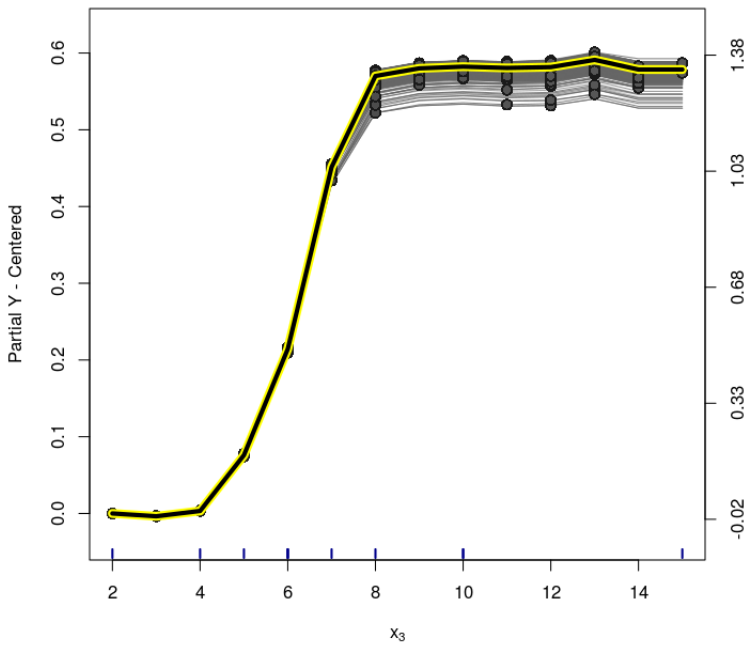

Figure C.62: *PS-BART model (Simulation Based on Real Data) - Variable x_3 - centered-ICE Plot for the treatment effect.*

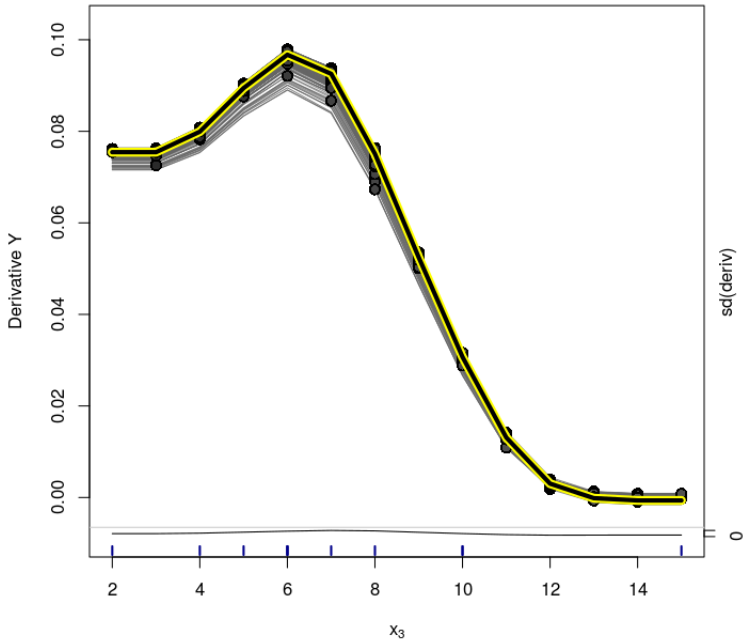

Figure C.63: *PS-BART model (Simulation Based on Real Data) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

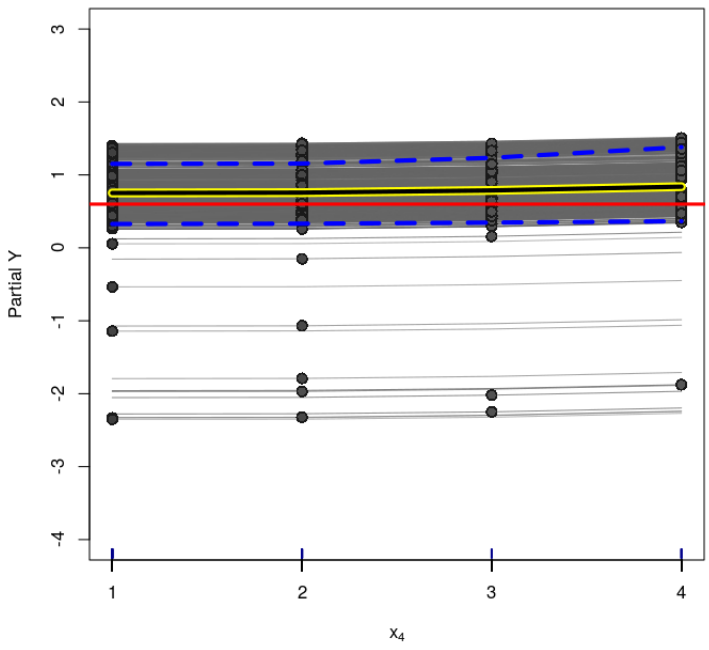

Figure C.64: *PS-BART model (Simulation Based on Real Data) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

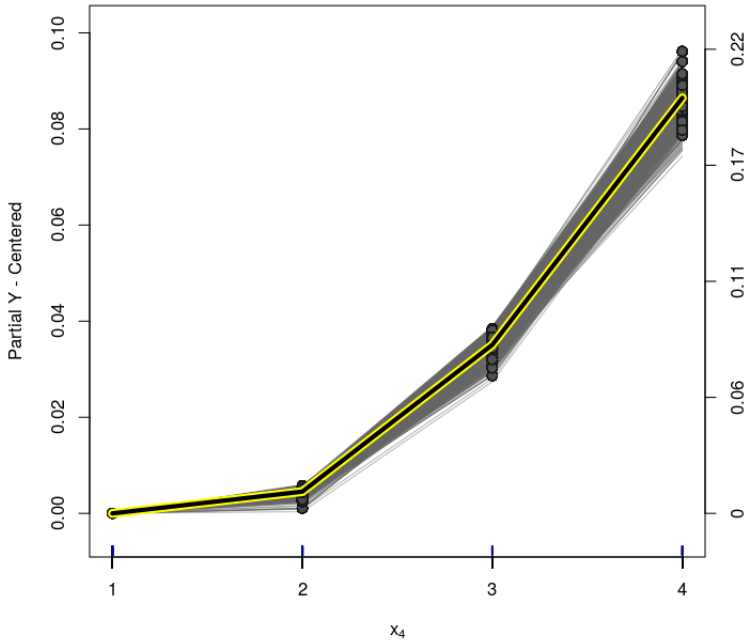

Figure C.65: *PS-BART model (Simulation Based on Real Data) - Variable x_4 - centered-ICE Plot for the treatment effect.*

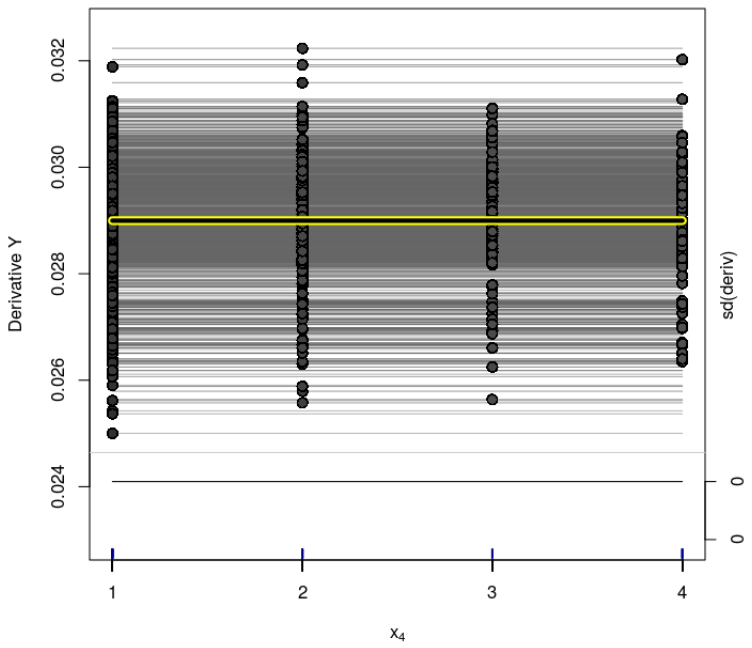

Figure C.66: *PS-BART model (Simulation Based on Real Data) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

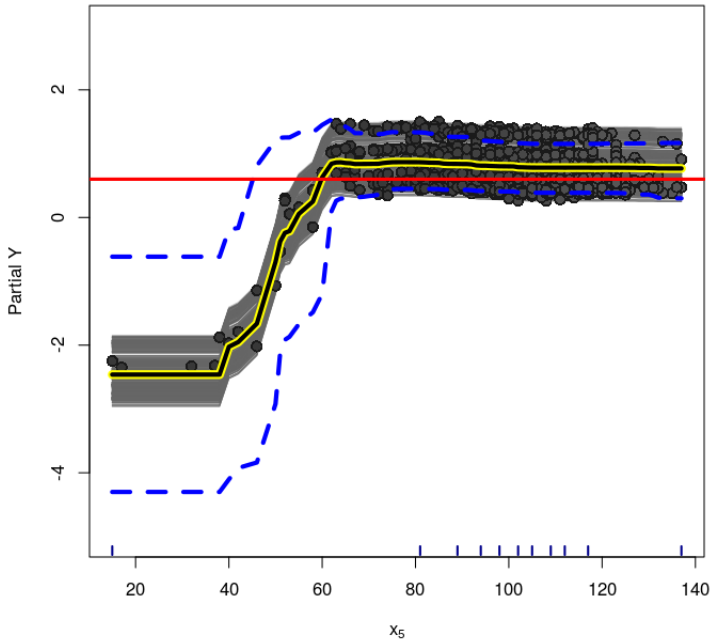

Figure C.67: *PS-BART model (Simulation Based on Real Data) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

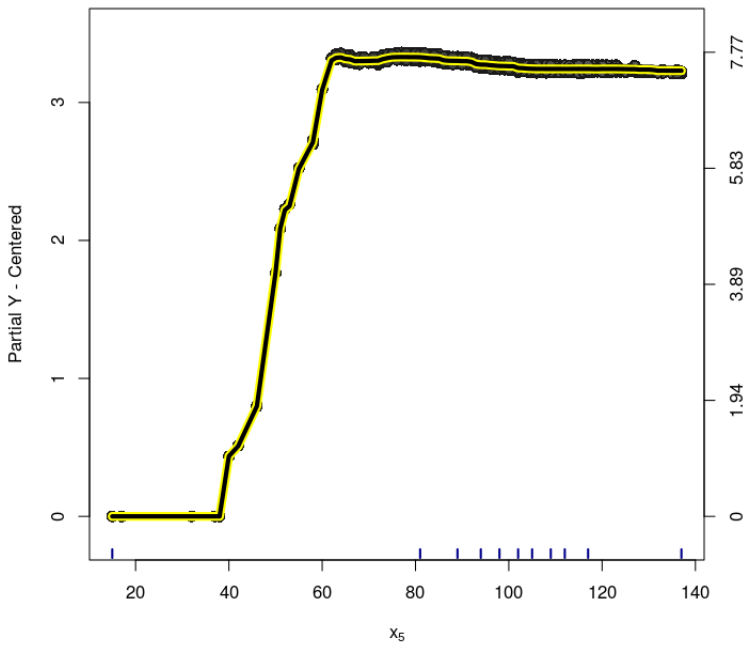

Figure C.68: *PS-BART model (Simulation Based on Real Data) - Variable x_5 - centered-ICE Plot for the treatment effect.*

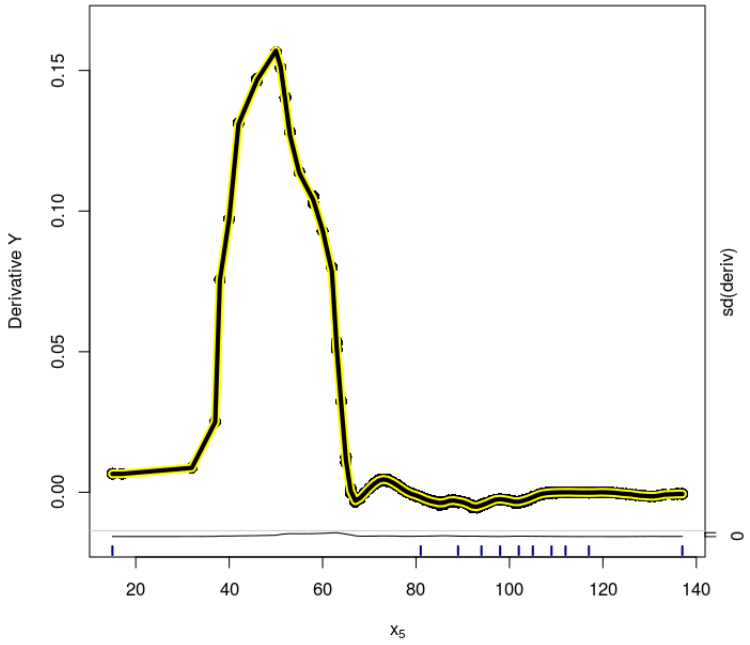

Figure C.69: *PS-BART model (Simulation Based on Real Data) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

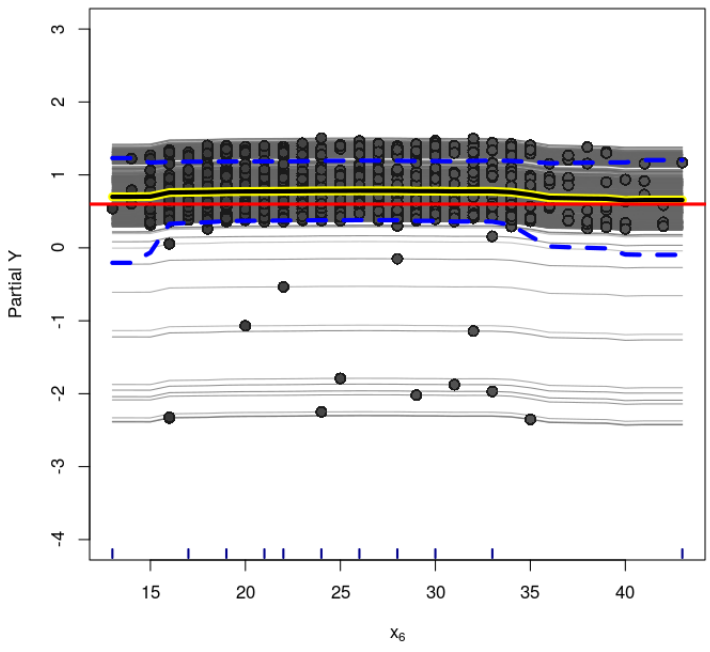

Figure C.70: *PS-BART model (Simulation Based on Real Data) - Variable x_6 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

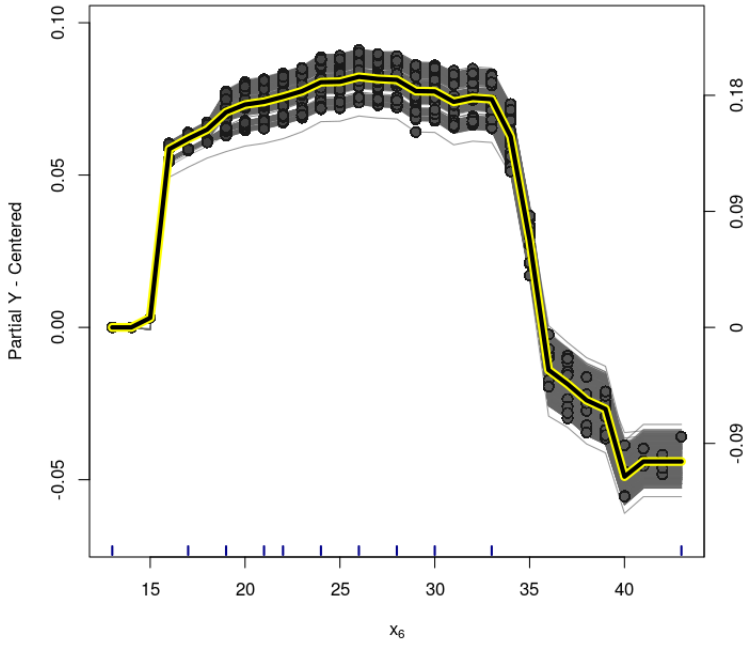

Figure C.71: *PS-BART model (Simulation Based on Real Data) - Variable x_6 - centered-ICE Plot for the treatment effect.*

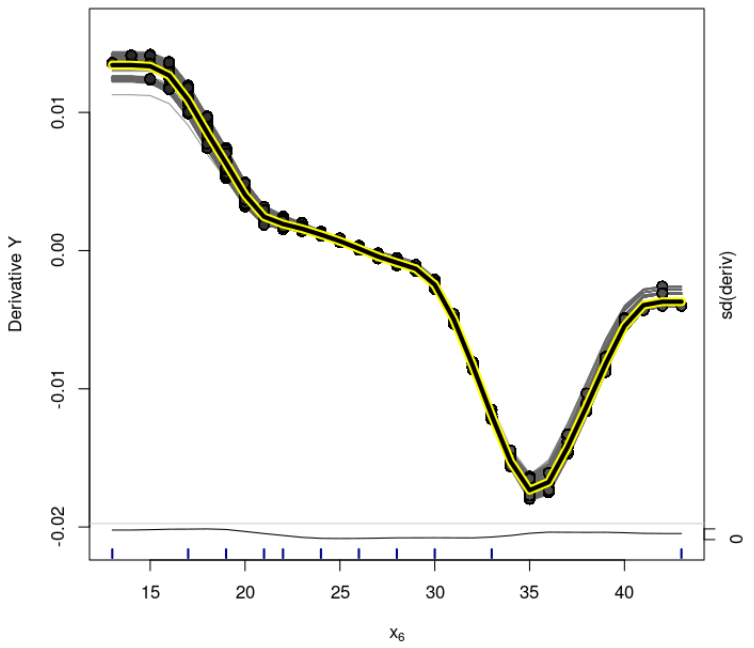

Figure C.72: *PS-BART model (Simulation Based on Real Data) - Variable x_6 - d-ICE Plot for the treatment effect estimates.*

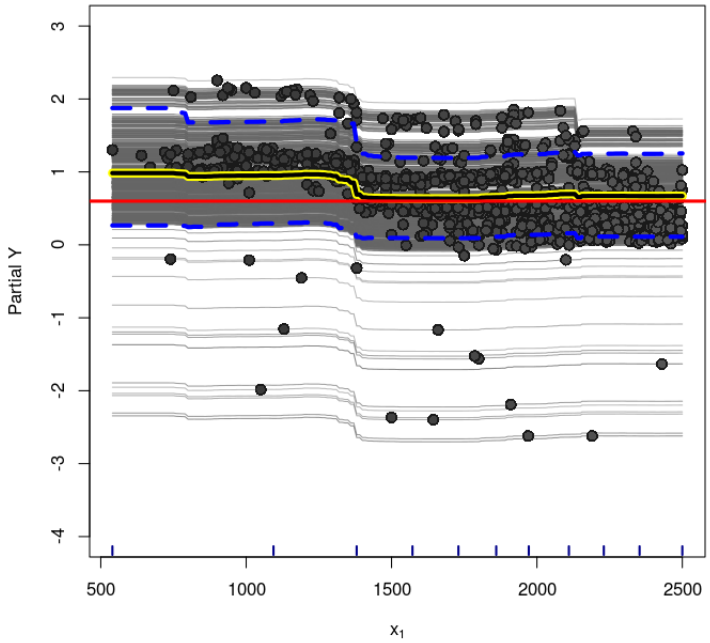

Figure C.73: *GLM-BART model (Simulation Based on Real Data) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

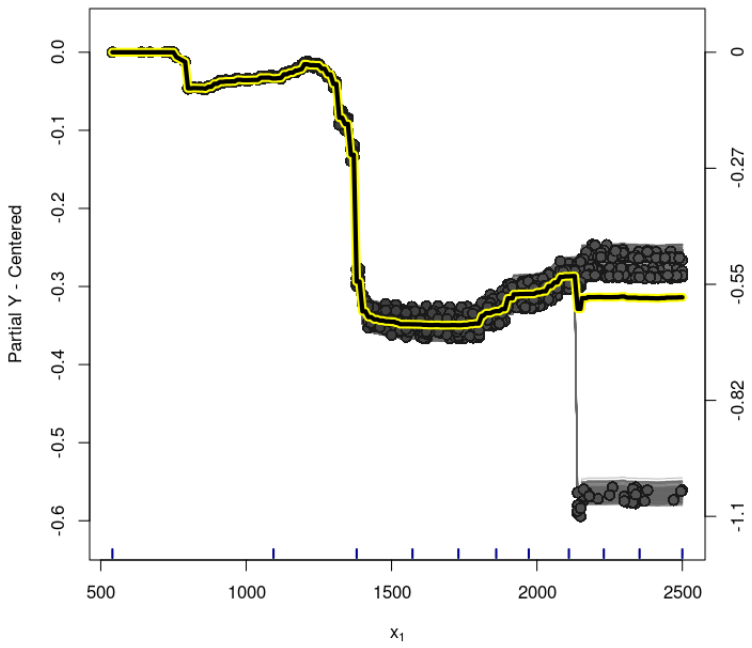

Figure C.74: *GLM-BART model (Simulation Based on Real Data) - Variable x_1 - centered-ICE Plot for the treatment effect.*

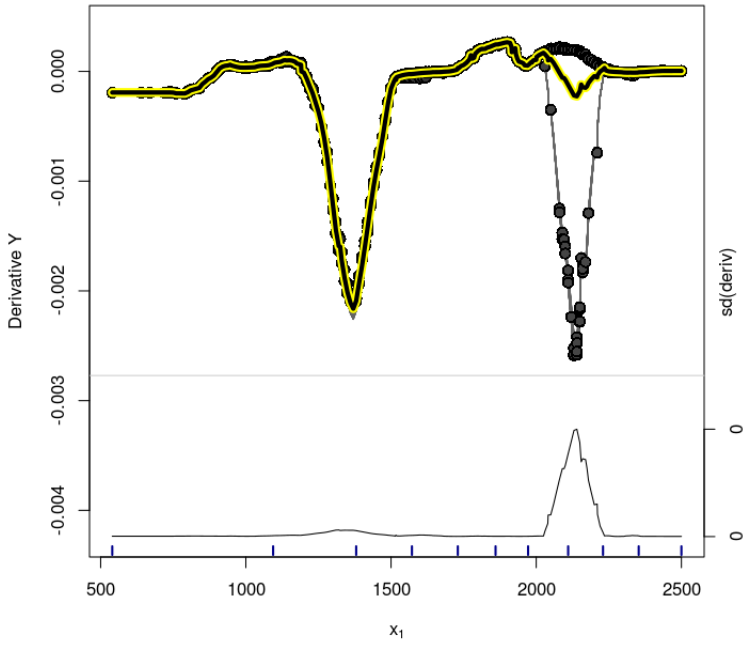

Figure C.75: *GLM-BART model (Simulation Based on Real Data) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

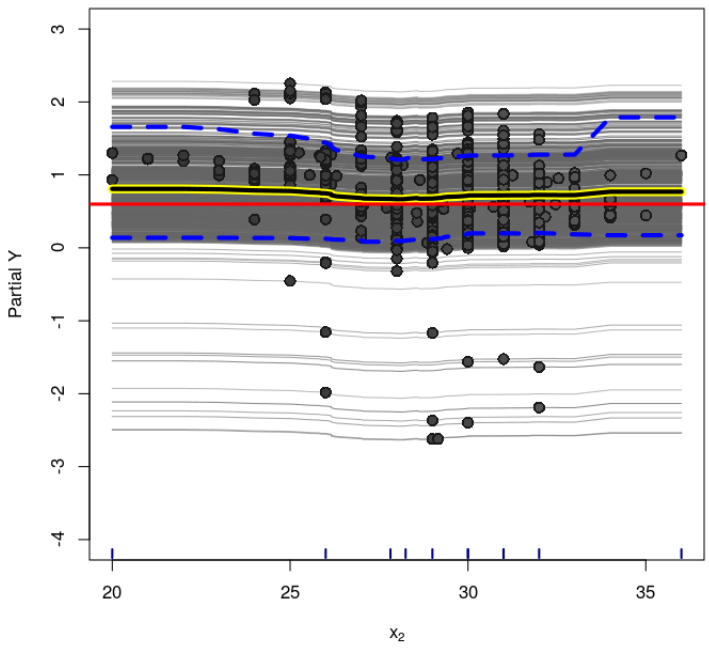

Figure C.76: *GLM-BART model (Simulation Based on Real Data) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

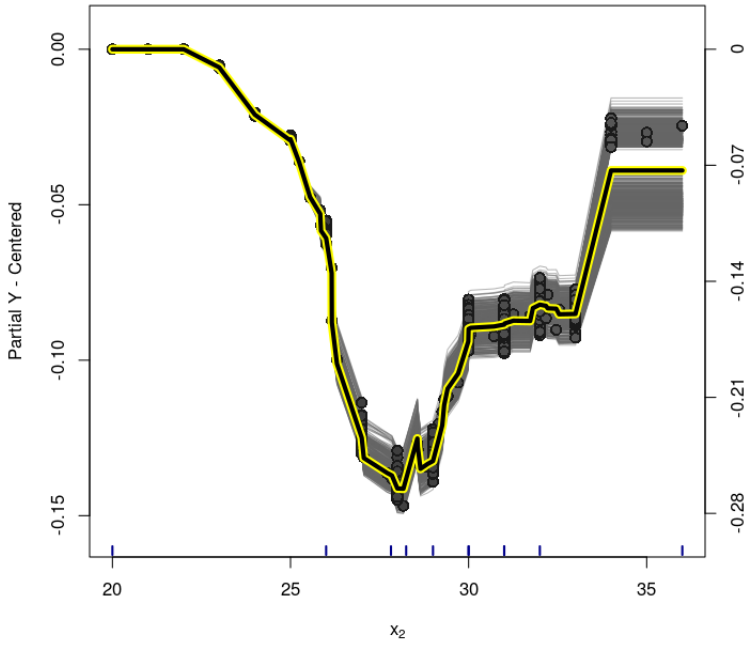

Figure C.77: *GLM-BART model (Simulation Based on Real Data) - Variable x_2 - centered-ICE Plot for the treatment effect.*

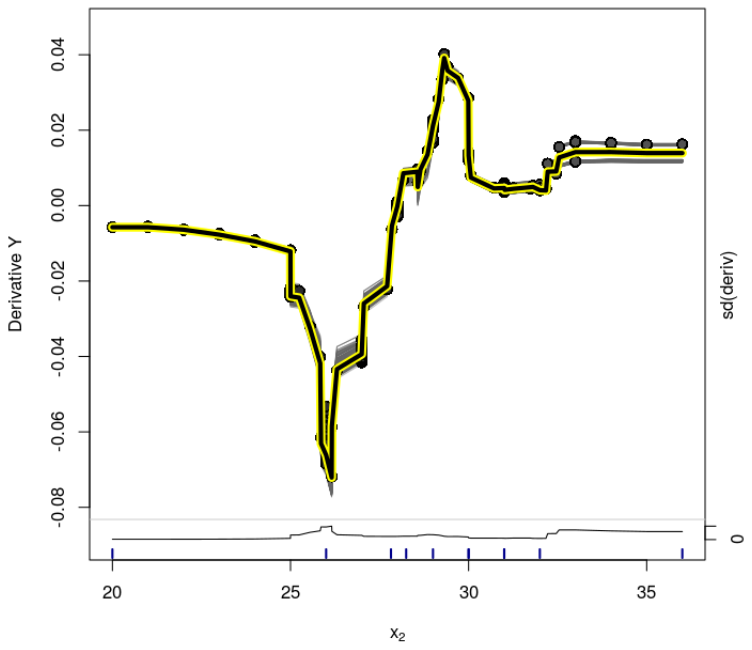

Figure C.78: *GLM-BART model (Simulation Based on Real Data) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

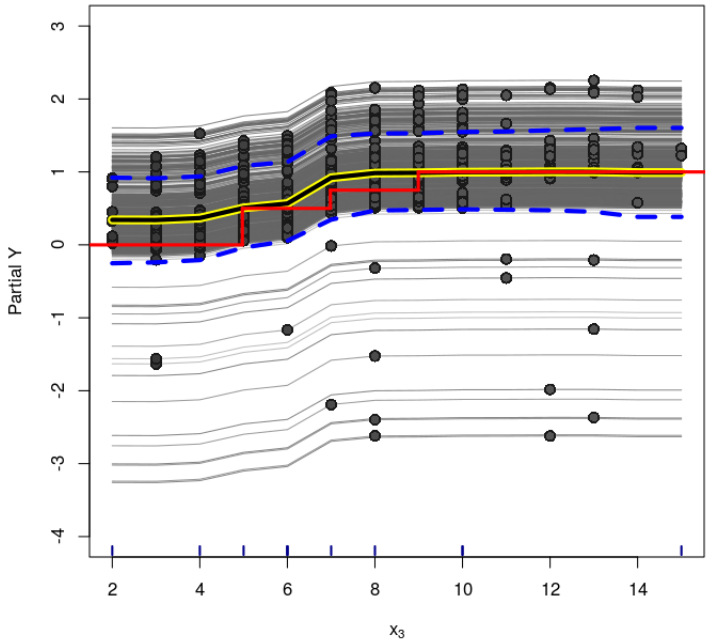

Figure C.79: *GLM-BART model (Simulation Based on Real Data) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

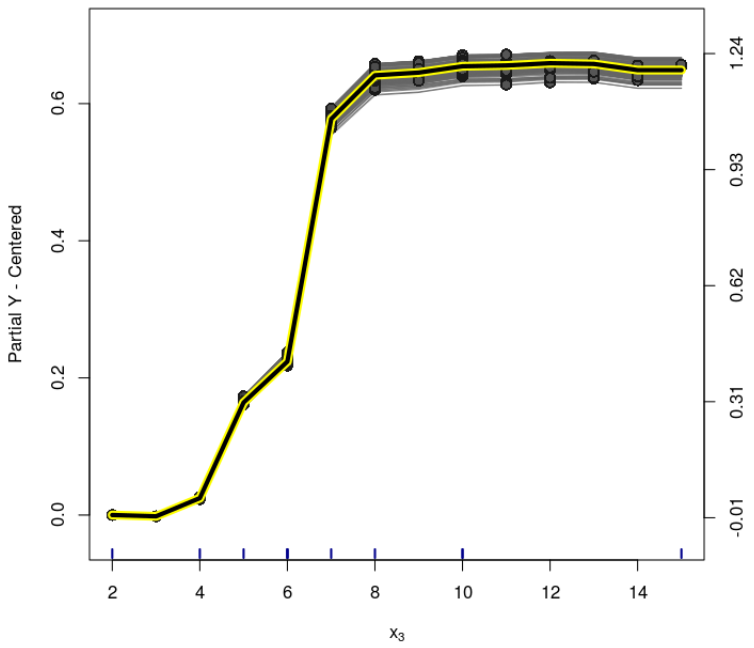

Figure C.80: *GLM-BART model (Simulation Based on Real Data) - Variable x_3 - centered-ICE Plot for the treatment effect.*

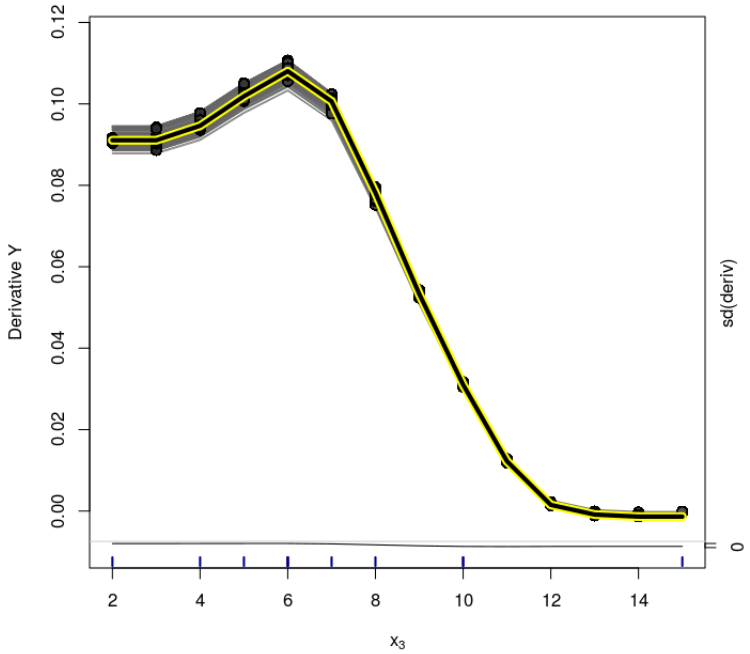

Figure C.81: *GLM-BART model (Simulation Based on Real Data) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

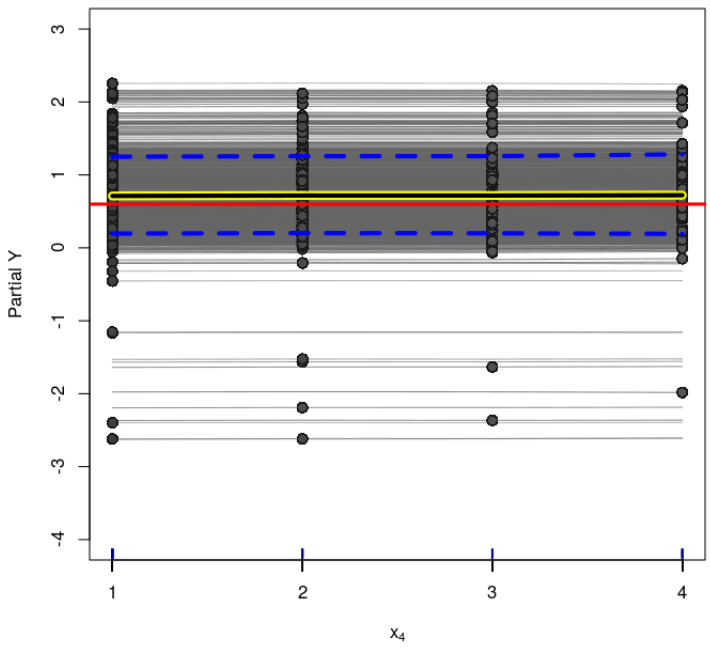

Figure C.82: *GLM-BART model (Simulation Based on Real Data) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

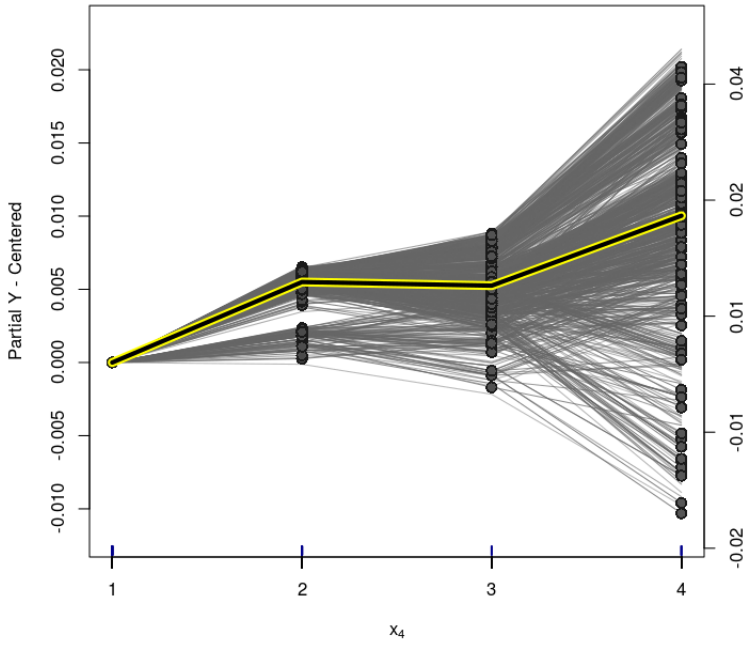

Figure C.83: *GLM-BART model (Simulation Based on Real Data) - Variable x_4 - centered-ICE Plot for the treatment effect.*

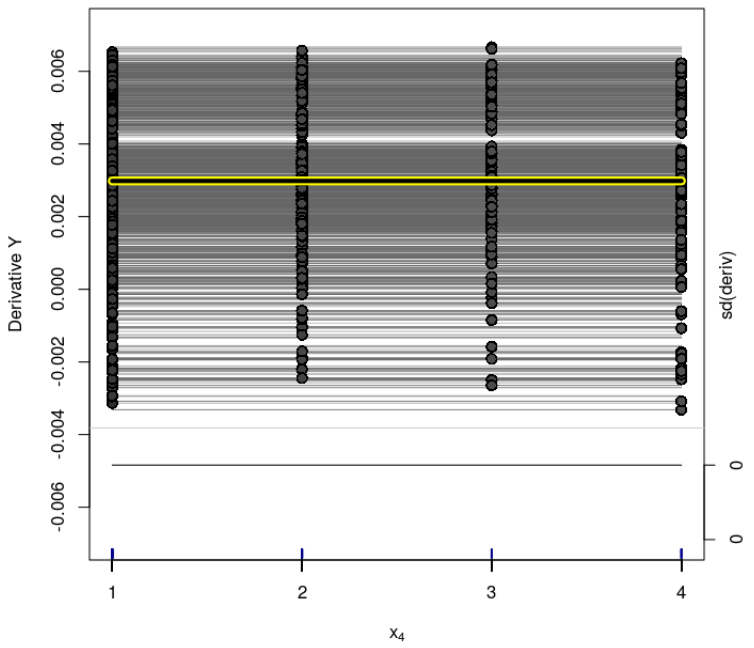

Figure C.84: *GLM-BART model (Simulation Based on Real Data) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

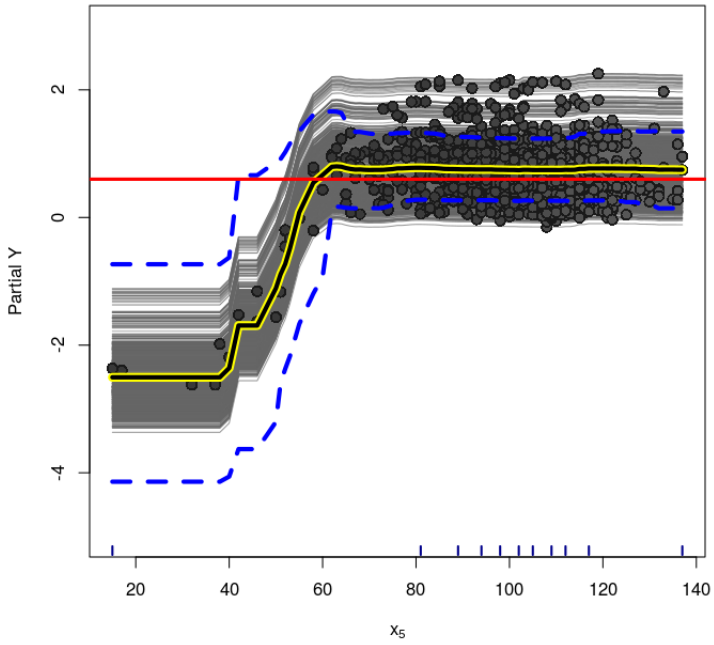

Figure C.85: *GLM-BART model (Simulation Based on Real Data) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

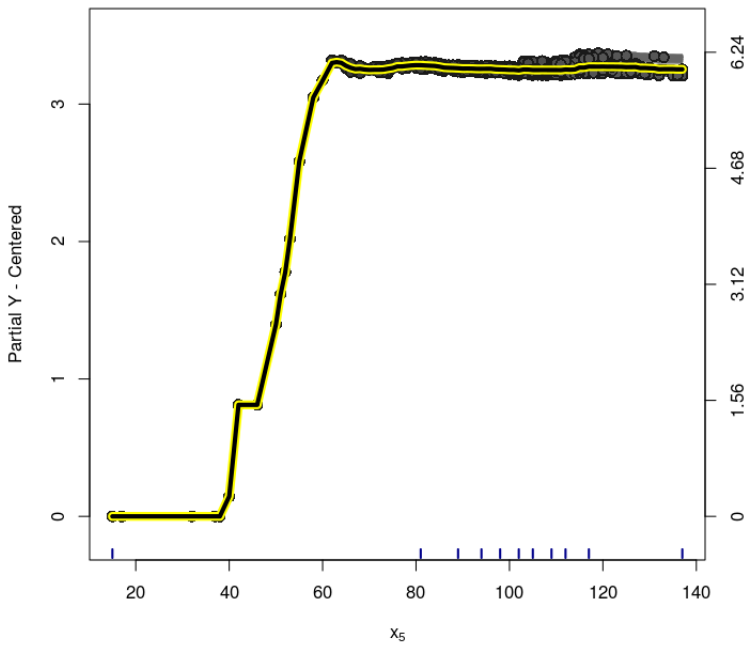

Figure C.86: *GLM-BART model (Simulation Based on Real Data) - Variable x_5 - centered-ICE Plot for the treatment effect.*

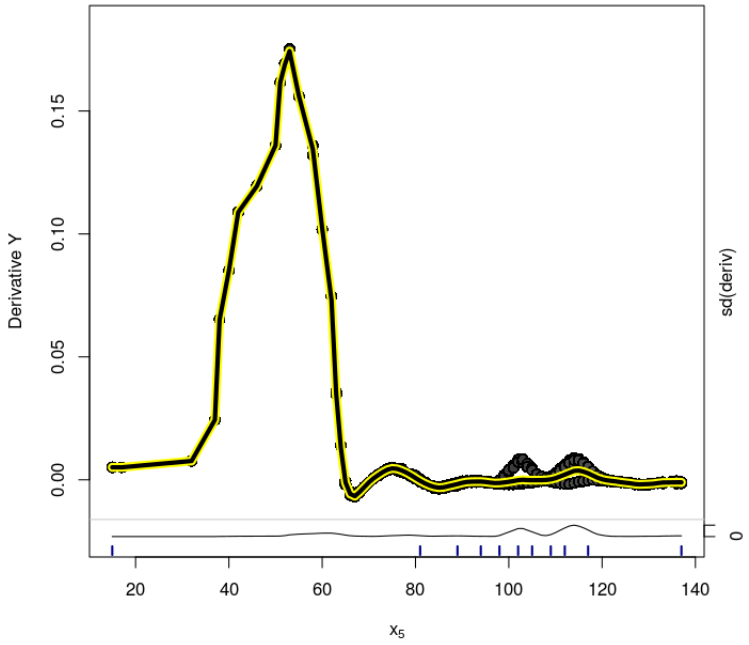

Figure C.87: *GLM-BART model (Simulation Based on Real Data) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

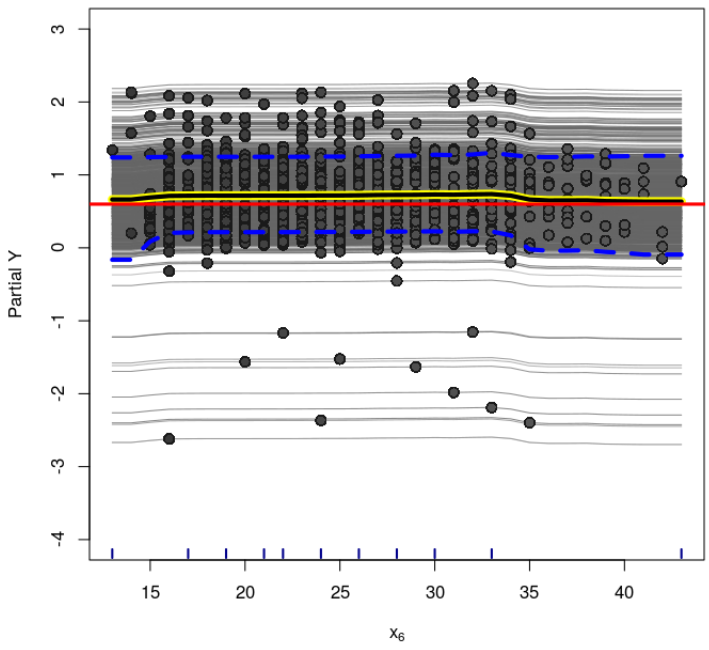

Figure C.88: *GLM-BART model (Simulation Based on Real Data) - Variable x_6 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

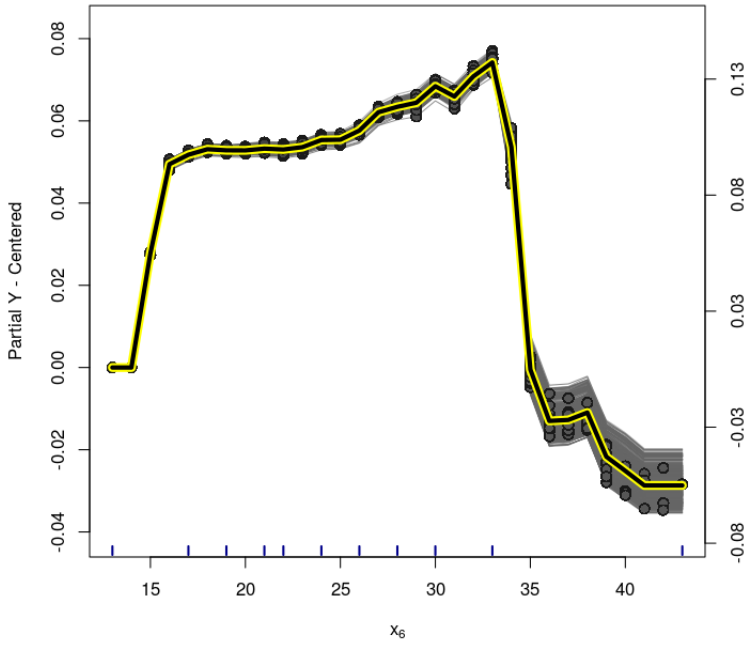

Figure C.89: *GLM-BART model (Simulation Based on Real Data) - Variable x_6 - centered-ICE Plot for the treatment effect.*

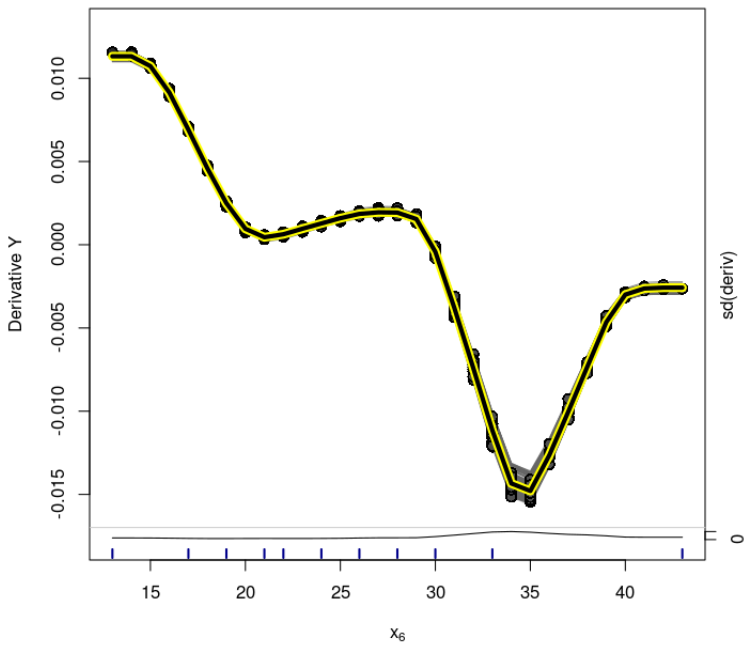

Figure C.90: *GLM-BART model (Simulation Based on Real Data) - Variable x_6 - d-ICE Plot for the treatment effect estimates.*

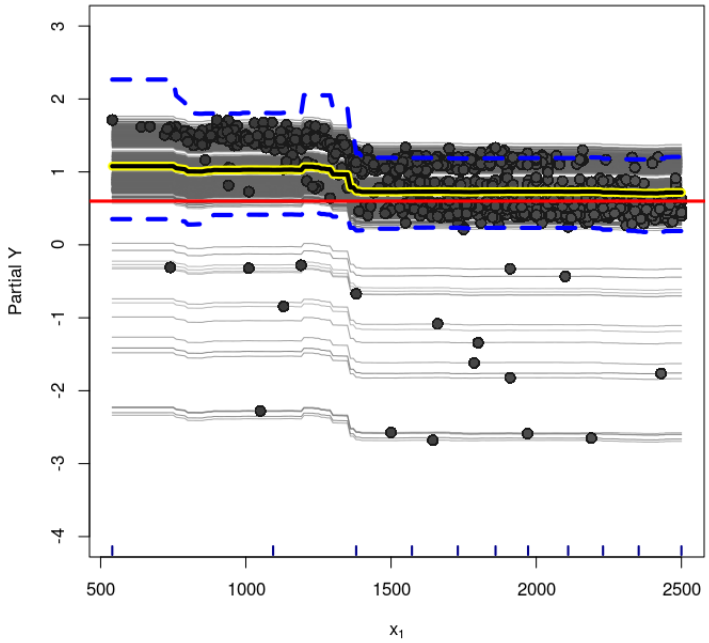

Figure C.91: *Rand-BART model (Simulation Based on Real Data) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

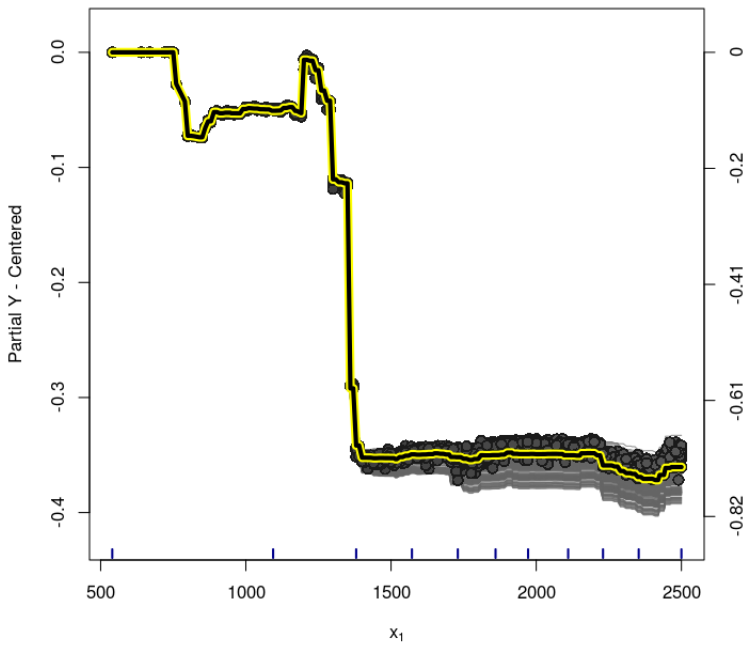

Figure C.92: *Rand-BART model (Simulation Based on Real Data) - Variable x_1 - centered-ICE Plot for the treatment effect.*

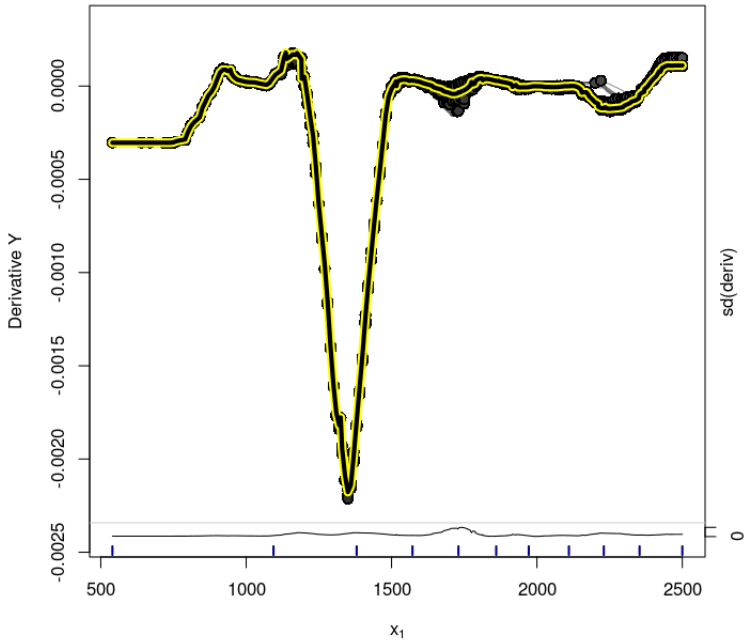

Figure C.93: *Rand-BART model (Simulation Based on Real Data) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

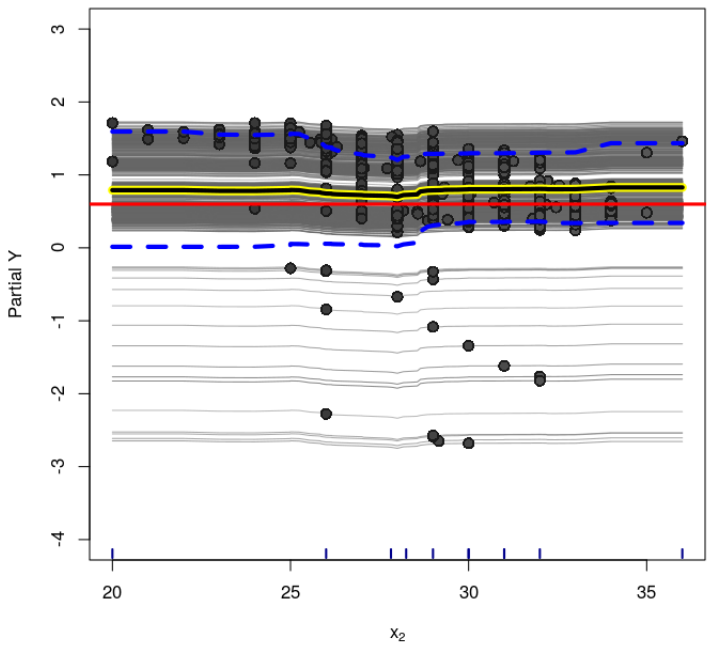

Figure C.94: *Rand-BART model (Simulation Based on Real Data) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

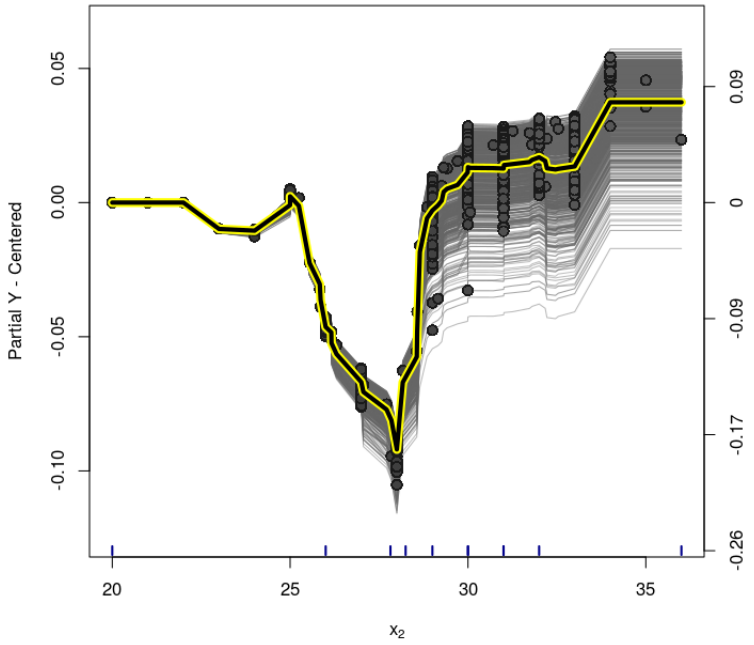

Figure C.95: *Rand-BART model (Simulation Based on Real Data) - Variable x_2 - centered-ICE Plot for the treatment effect.*

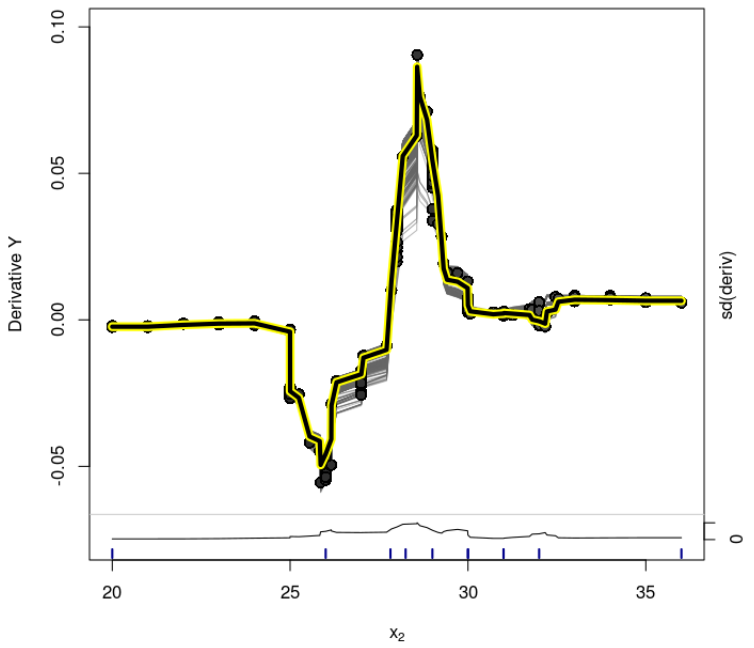

Figure C.96: *Rand-BART model (Simulation Based on Real Data) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

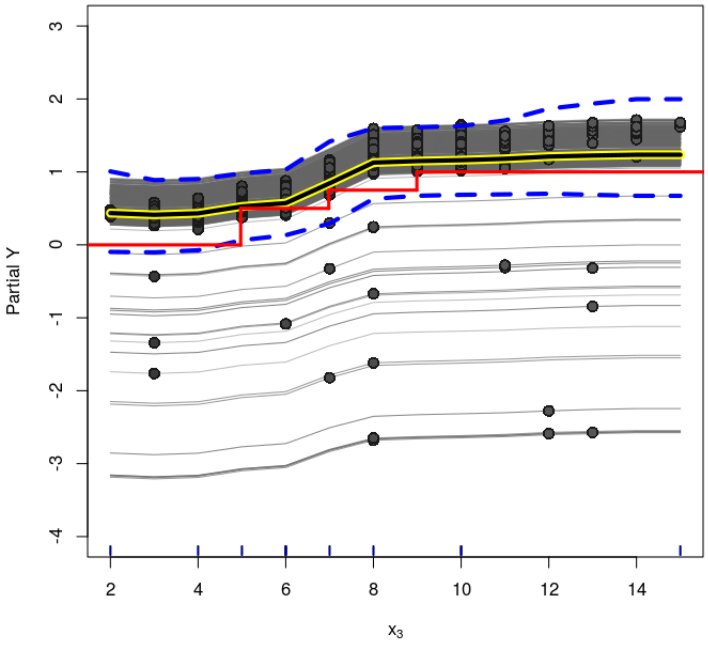

Figure C.97: *Rand-BART model (Simulation Based on Real Data) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

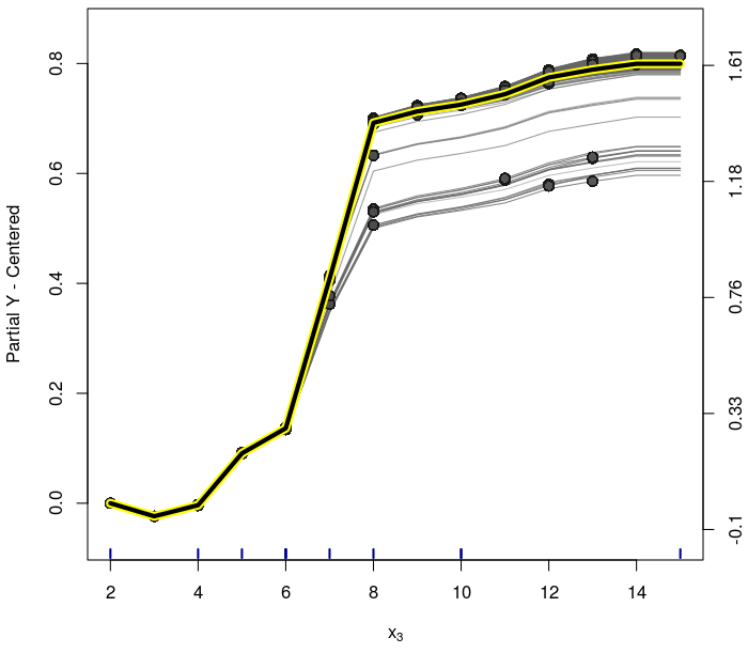

Figure C.98: *Rand-BART model (Simulation Based on Real Data) - Variable x_3 - centered-ICE Plot for the treatment effect.*

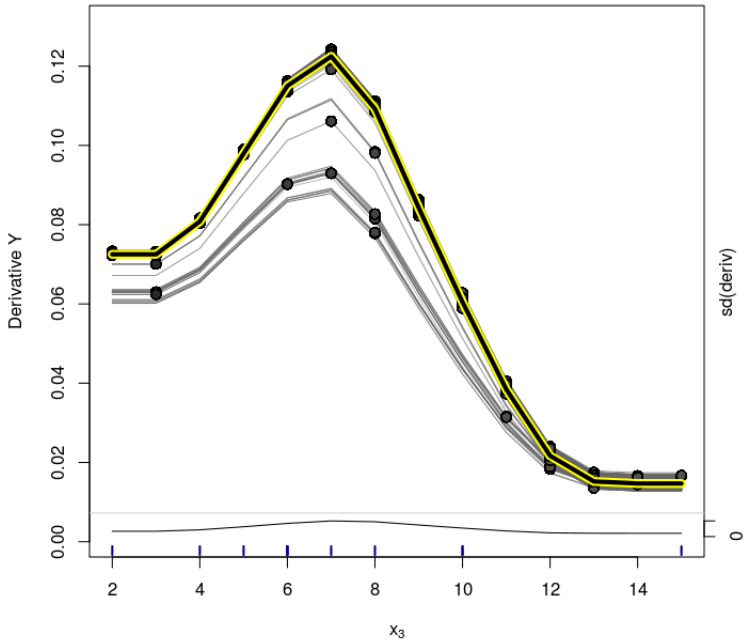

Figure C.99: *Rand-BART model (Simulation Based on Real Data) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

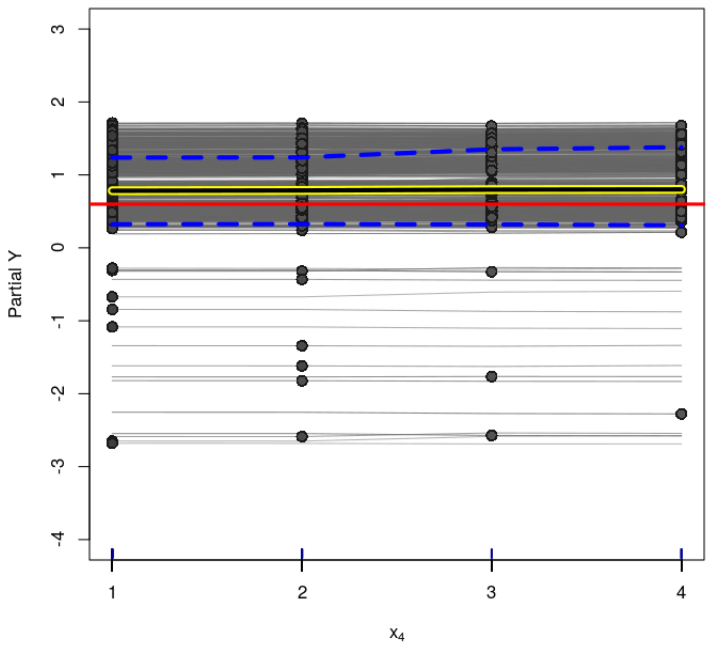

Figure C.100: *Rand-BART model (Simulation Based on Real Data) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

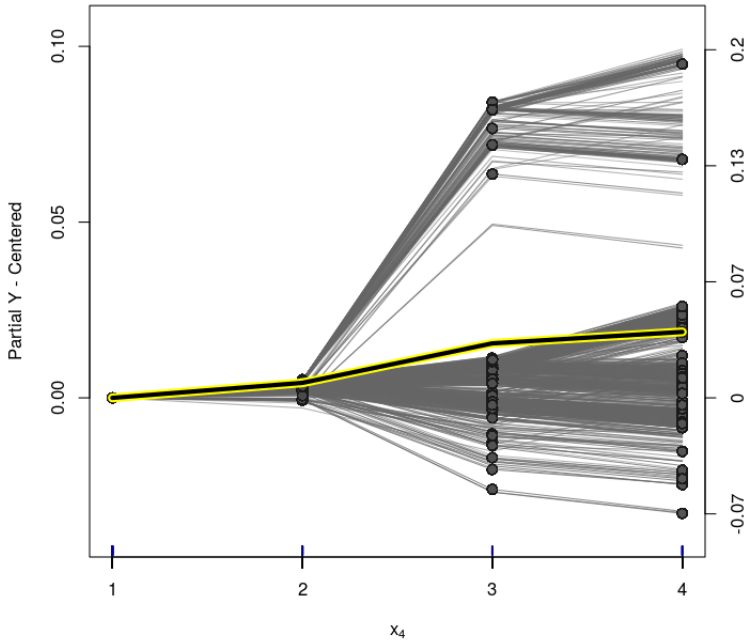

Figure C.101: *Rand-BART model (Simulation Based on Real Data) - Variable x_4 - centered-ICE Plot for the treatment effect.*

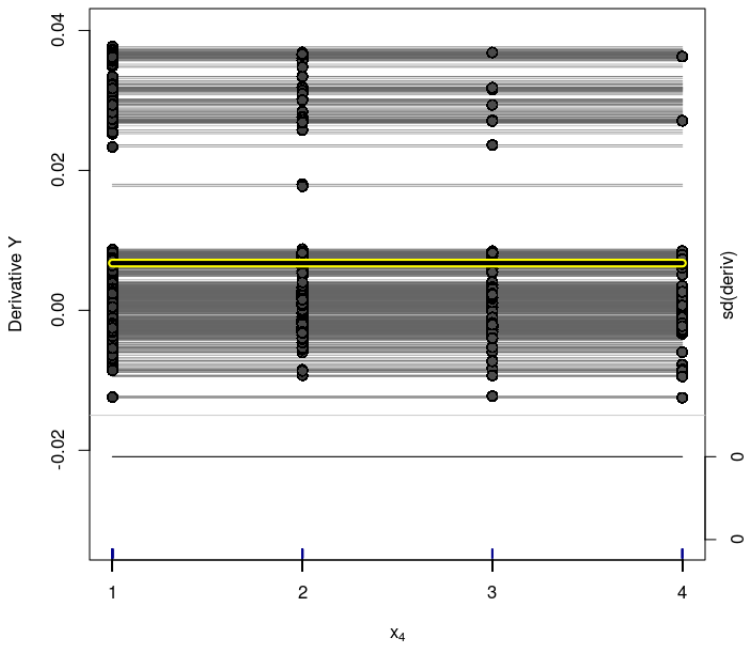

Figure C.102: *Rand-BART model (Simulation Based on Real Data) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

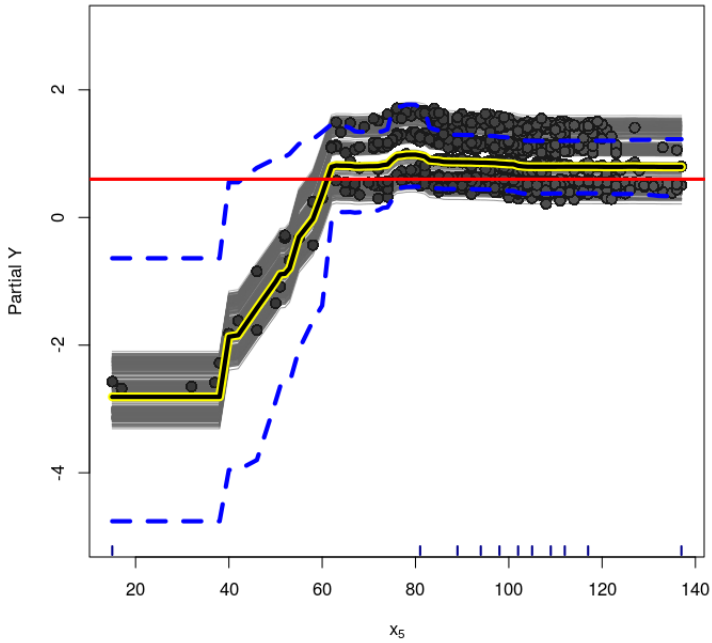

Figure C.103: *Rand-BART model (Simulation Based on Real Data) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

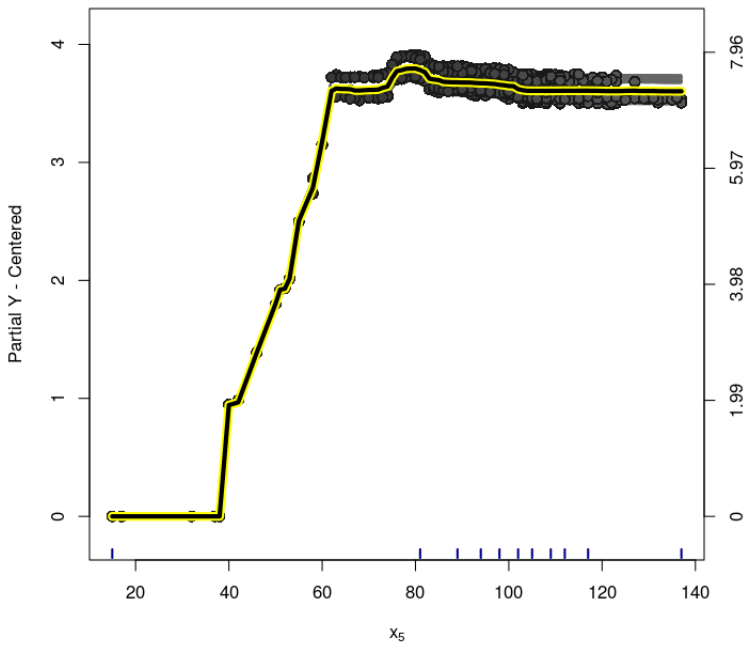

Figure C.104: *Rand-BART model (Simulation Based on Real Data) - Variable x_5 - centered-ICE Plot for the treatment effect.*

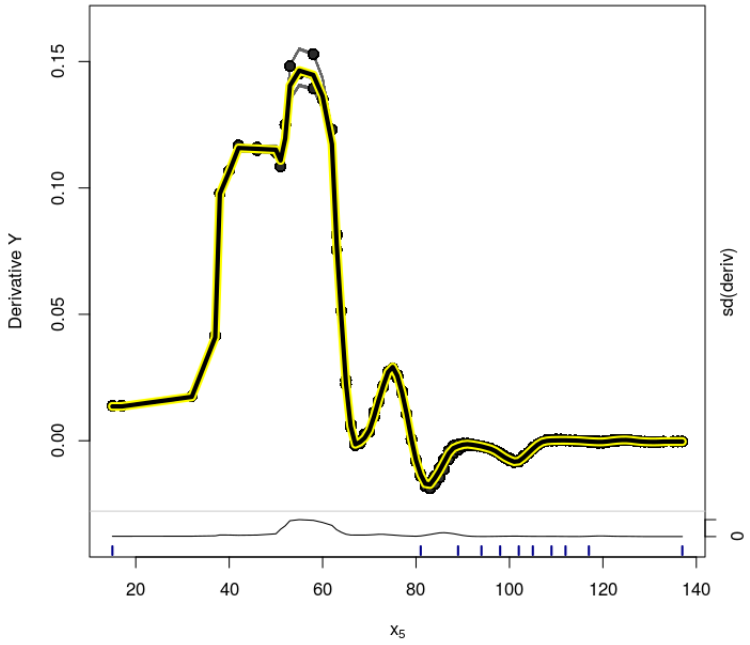

Figure C.105: *Rand-BART model (Simulation Based on Real Data) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

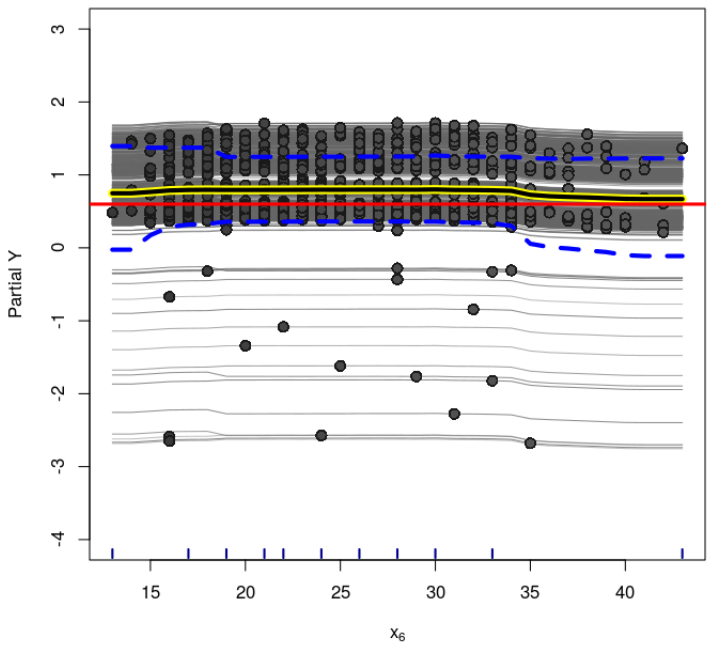

Figure C.106: *Rand-BART model (Simulation Based on Real Data) - Variable x_6 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

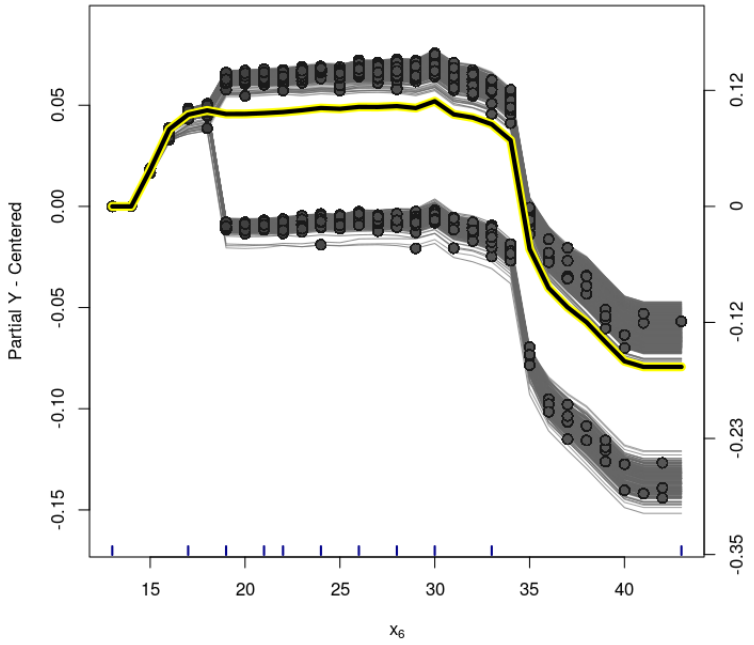

Figure C.107: *Rand-BART model (Simulation Based on Real Data) - Variable x_6 - centered-ICE Plot for the treatment effect.*

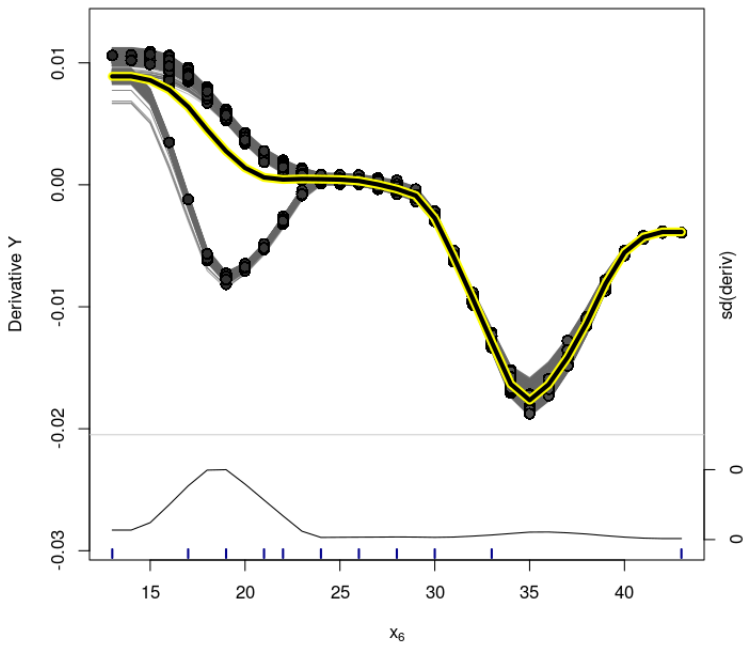

Figure C.108: *Rand-BART model (Simulation Based on Real Data) - Variable x_6 - d-ICE Plot for the treatment effect estimates.*

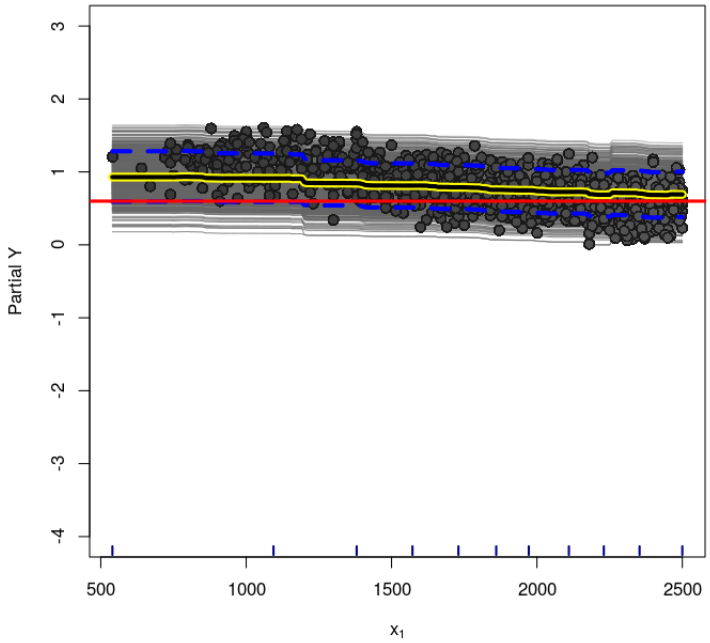

Figure C.109: *PS-BCF model (Simulation Based on Real Data) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

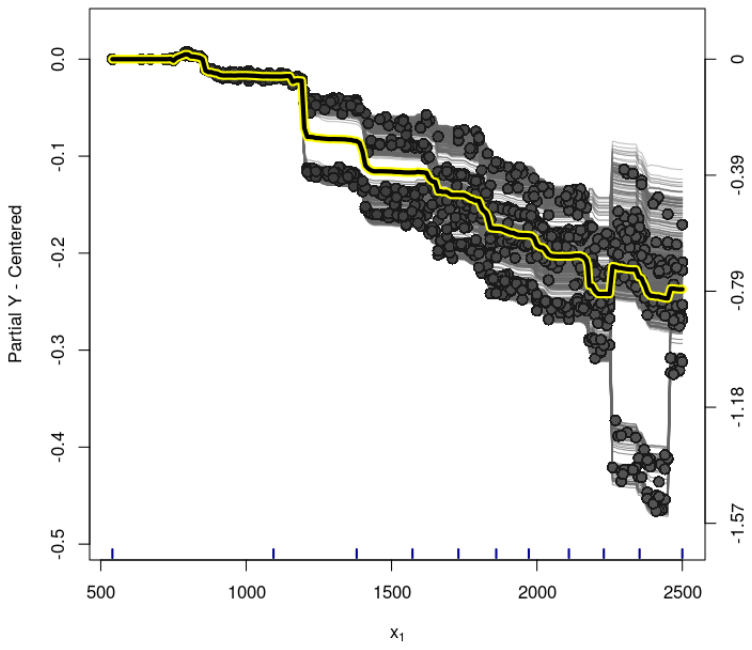

Figure C.110: *PS-BCF model (Simulation Based on Real Data) - Variable x_1 - centered-ICE Plot for the treatment effect.*

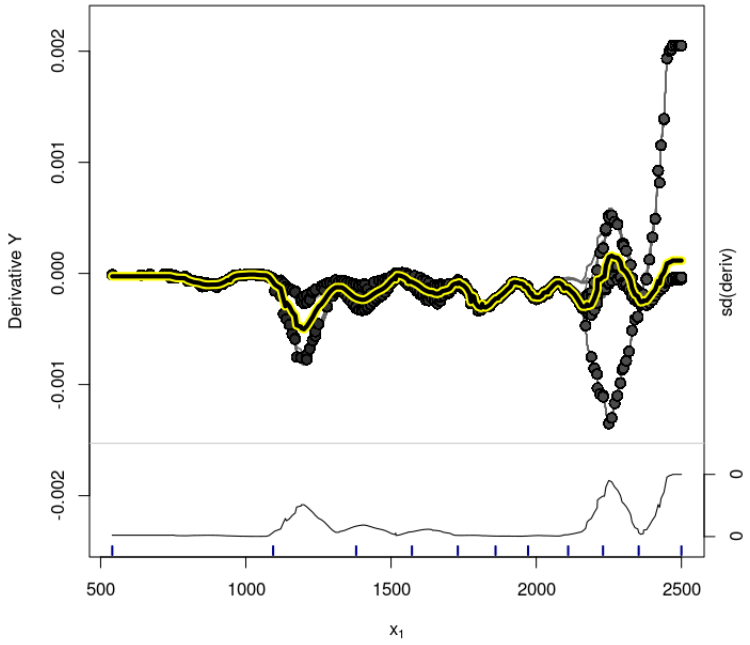

Figure C.111: *PS-BCF model (Simulation Based on Real Data) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

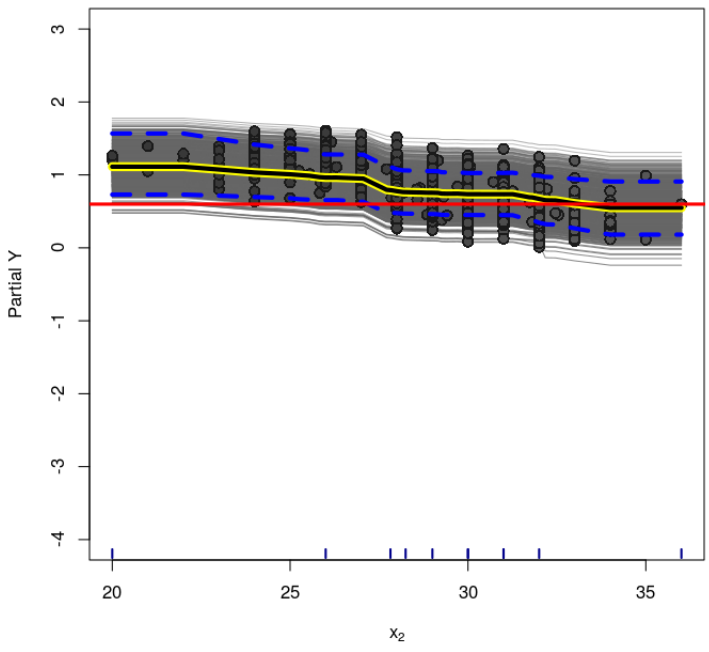

Figure C.112: *PS-BCF model (Simulation Based on Real Data) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

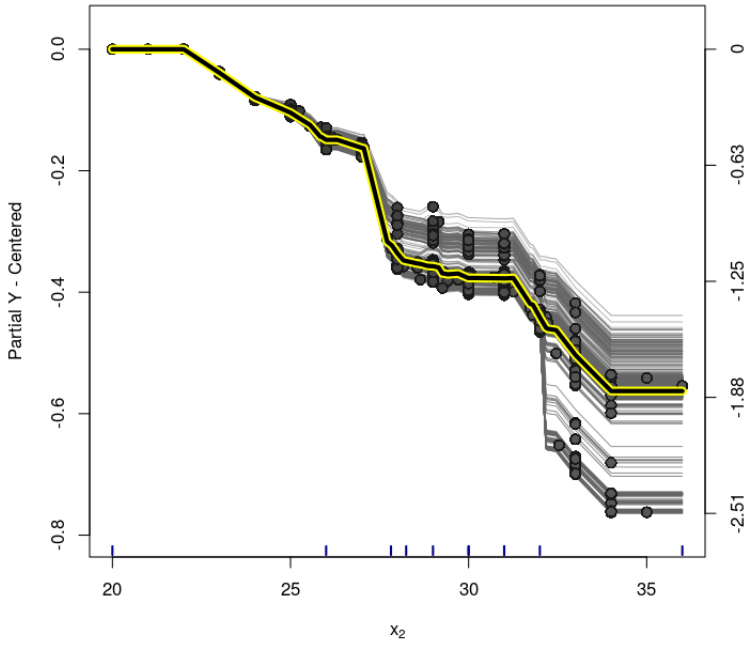

Figure C.113: *PS-BCF model (Simulation Based on Real Data) - Variable x_2 - centered-ICE Plot for the treatment effect.*

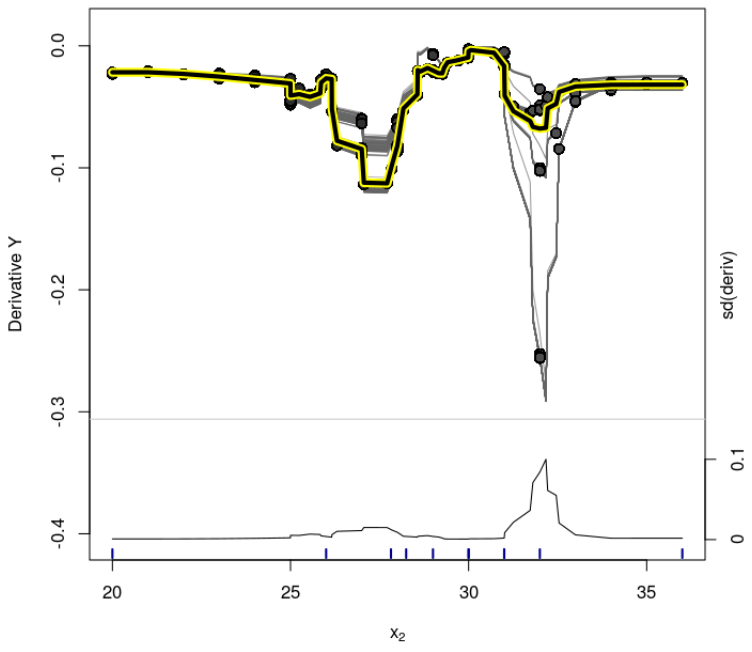

Figure C.114: *PS-BCF model (Simulation Based on Real Data) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

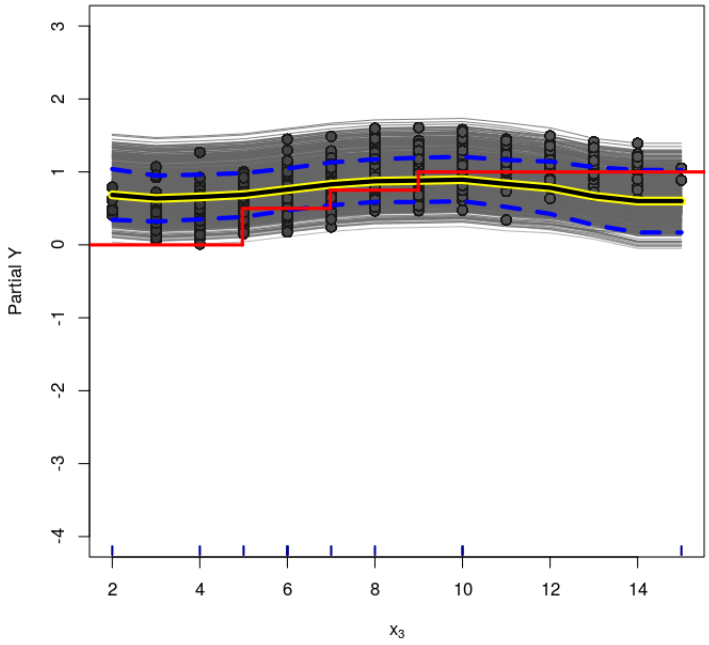

Figure C.115: *PS-BCF model (Simulation Based on Real Data) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

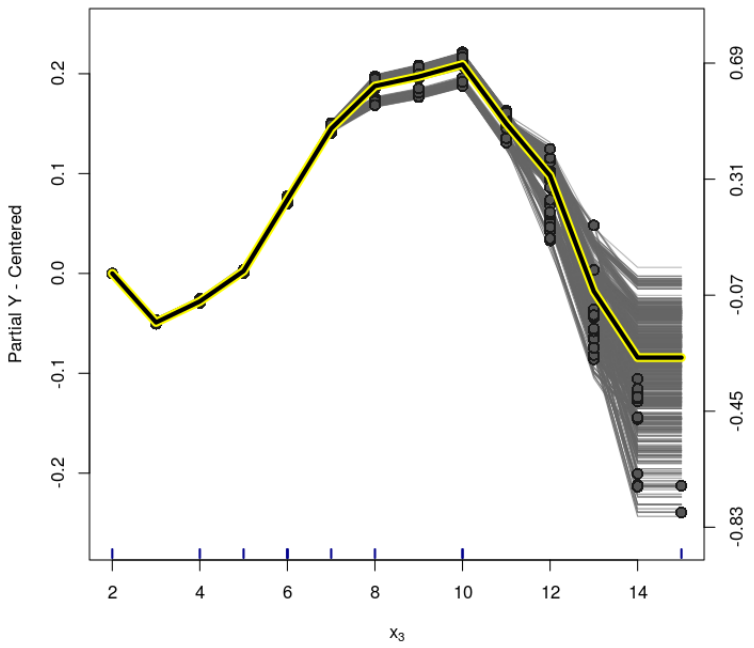

Figure C.116: *PS-BCF model (Simulation Based on Real Data) - Variable x_3 - centered-ICE Plot for the treatment effect.*

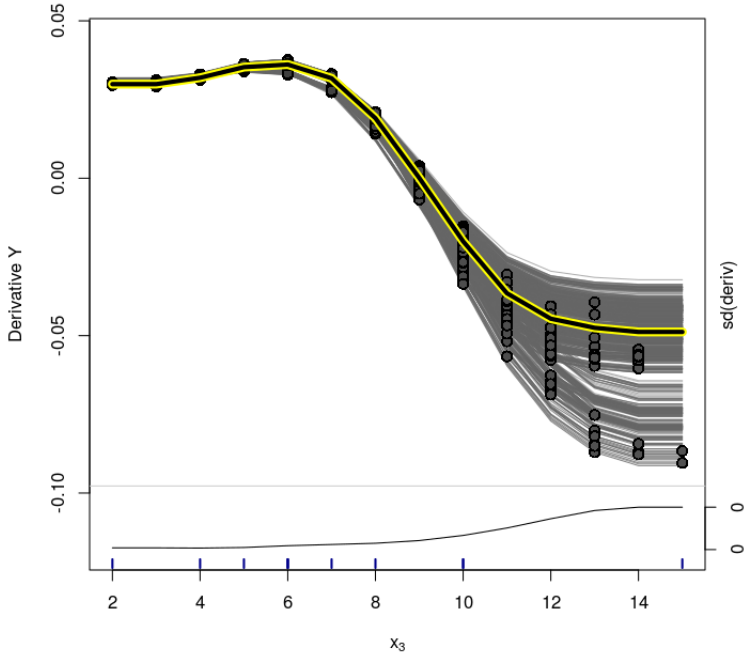

Figure C.117: *PS-BCF model (Simulation Based on Real Data) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

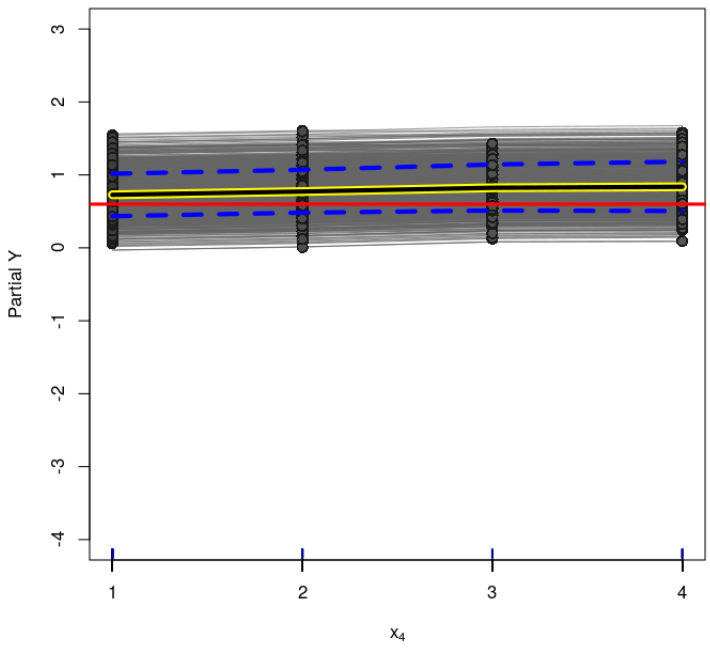

Figure C.118: *PS-BCF model (Simulation Based on Real Data) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

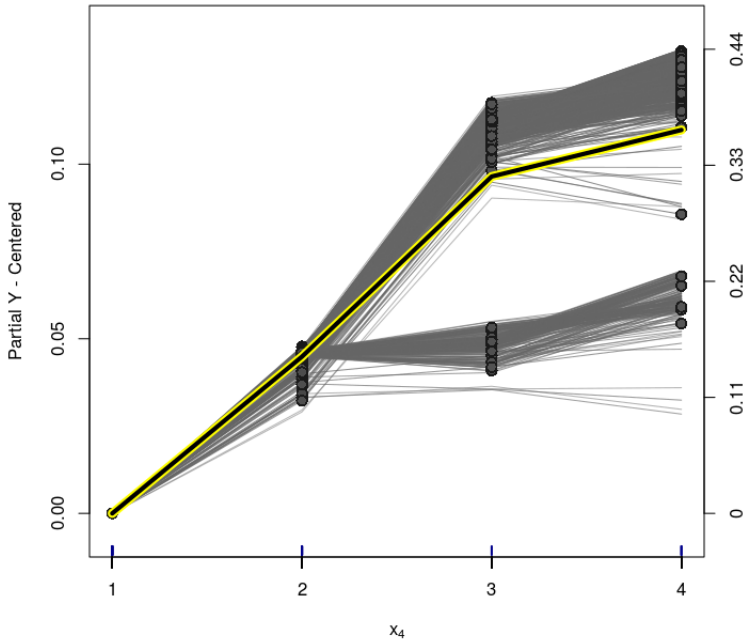

Figure C.119: *PS-BCF model (Simulation Based on Real Data) - Variable x_4 - centered-ICE Plot for the treatment effect.*

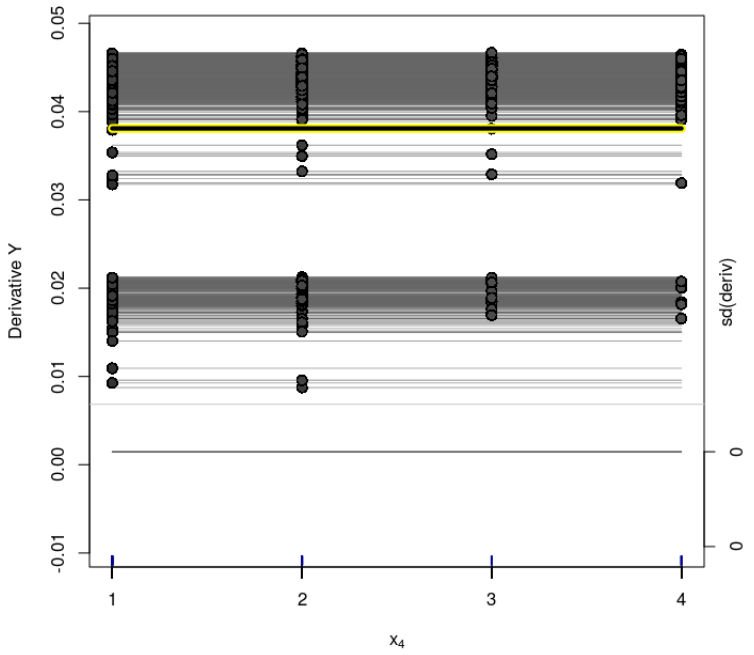

Figure C.120: *PS-BCF model (Simulation Based on Real Data) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

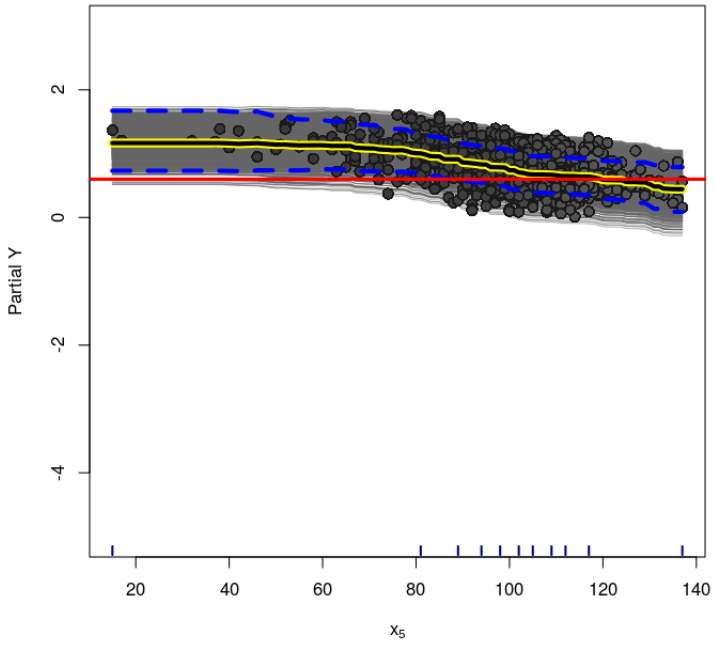

Figure C.121: *PS-BCF model (Simulation Based on Real Data) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

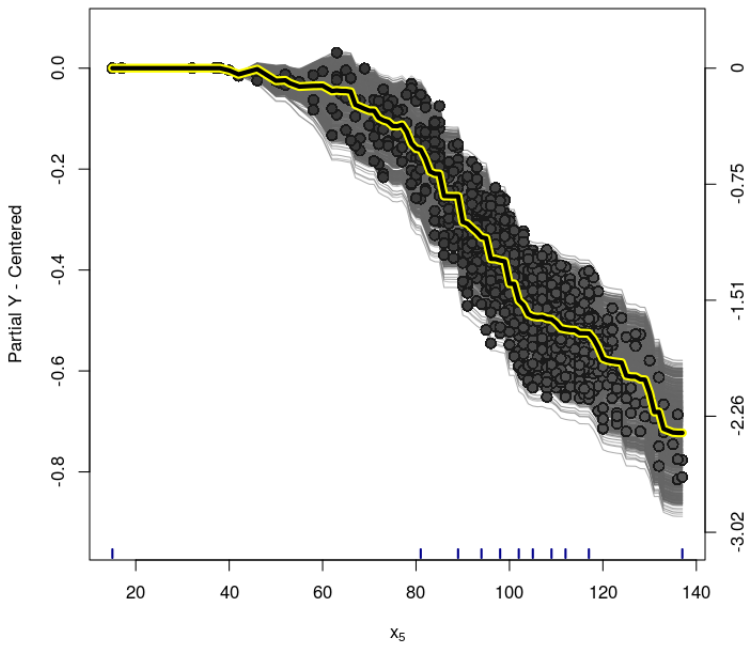

Figure C.122: *PS-BCF model (Simulation Based on Real Data) - Variable x_5 - centered-ICE Plot for the treatment effect.*

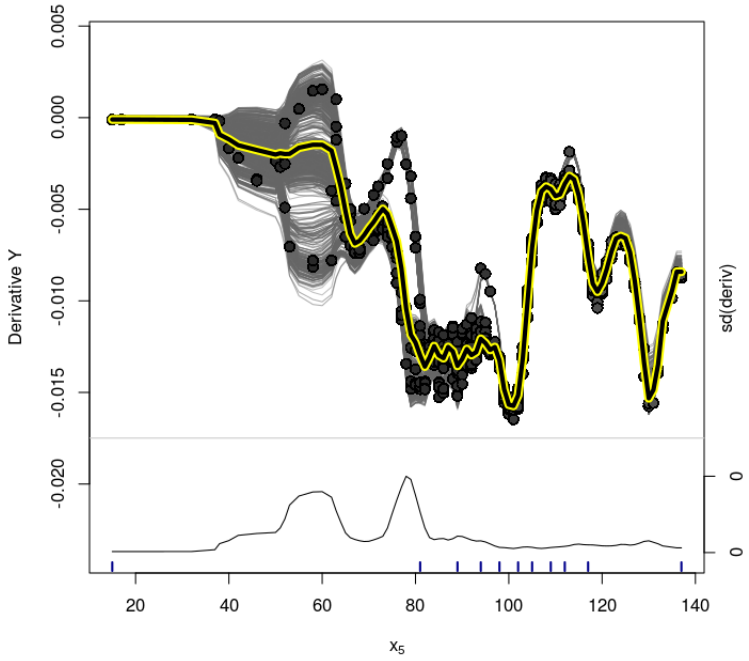

Figure C.123: *PS-BCF model (Simulation Based on Real Data) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

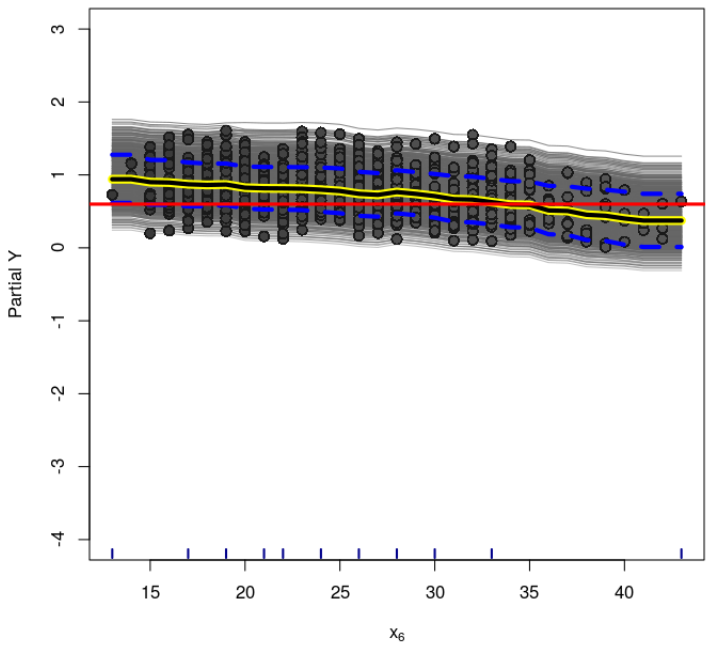

Figure C.124: *PS-BCF model (Simulation Based on Real Data) - Variable x_6 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

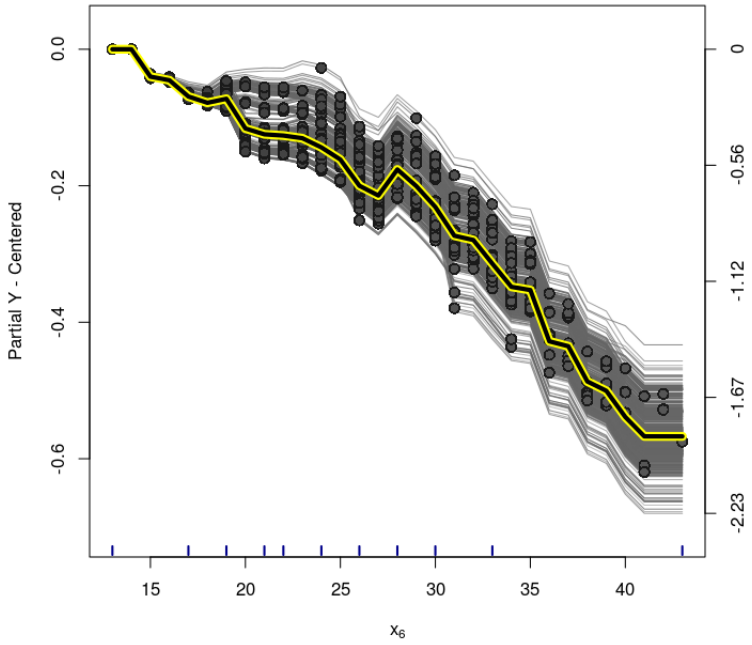

Figure C.125: *PS-BCF model (Simulation Based on Real Data) - Variable x_6 - centered-ICE Plot for the treatment effect.*

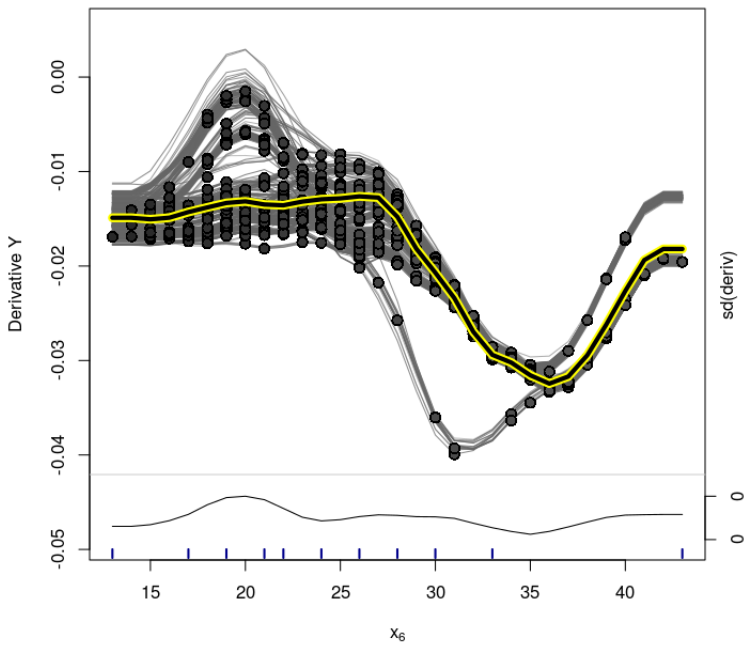

Figure C.126: *PS-BCF model (Simulation Based on Real Data) - Variable x_6 - d-ICE Plot for the treatment effect estimates.*

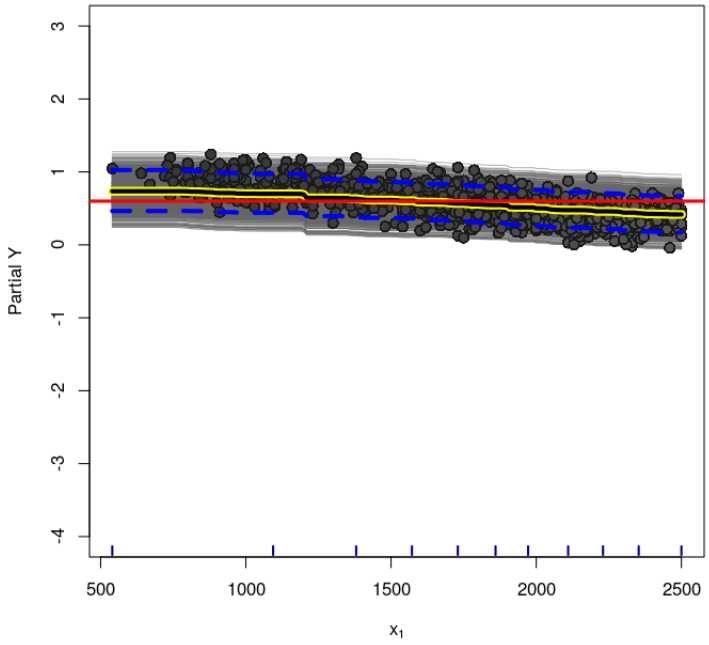

Figure C.127: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

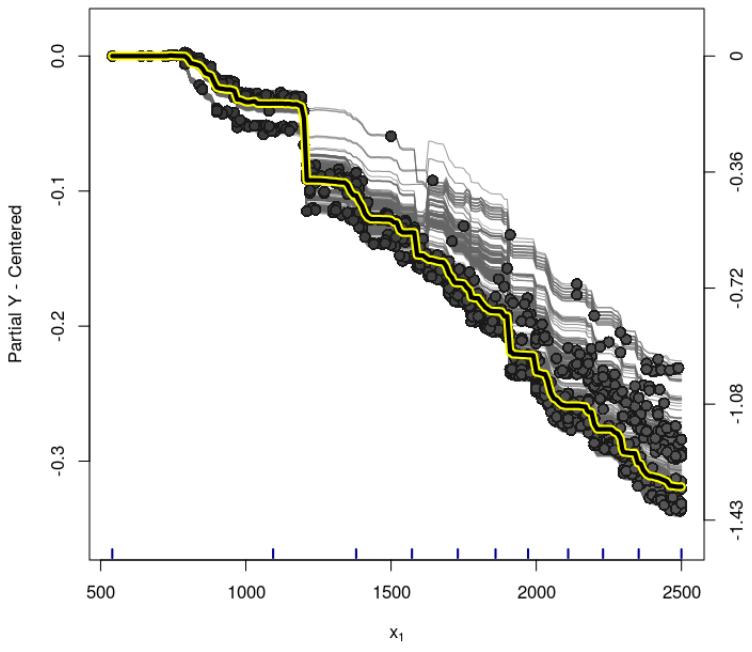

Figure C.128: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_1 - centered-ICE Plot for the treatment effect.*

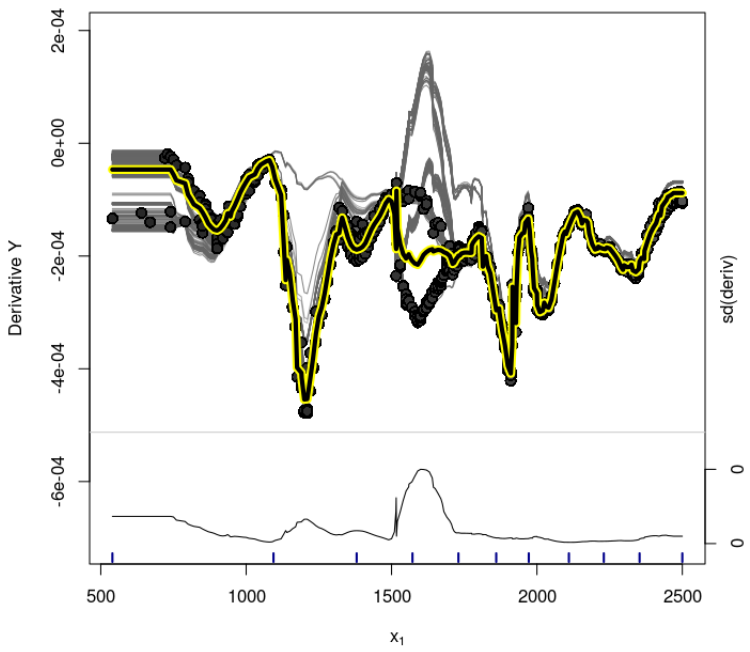

Figure C.129: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

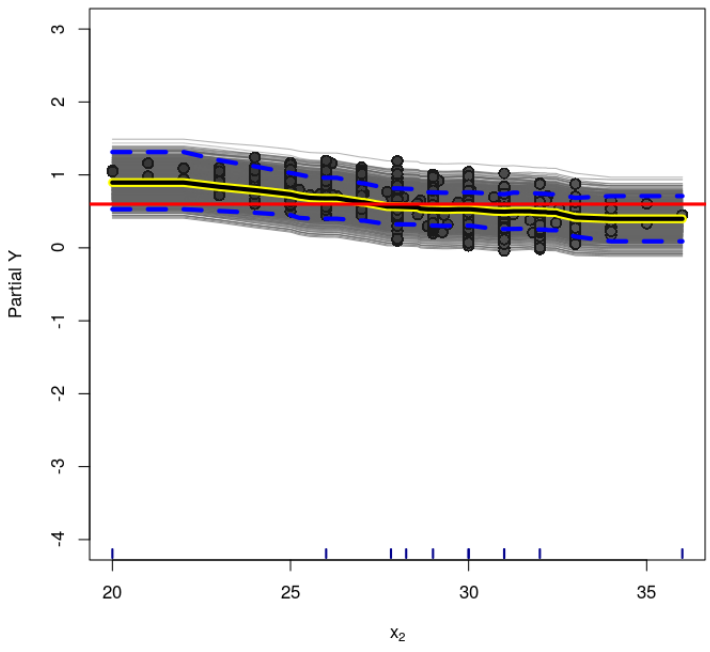

Figure C.130: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

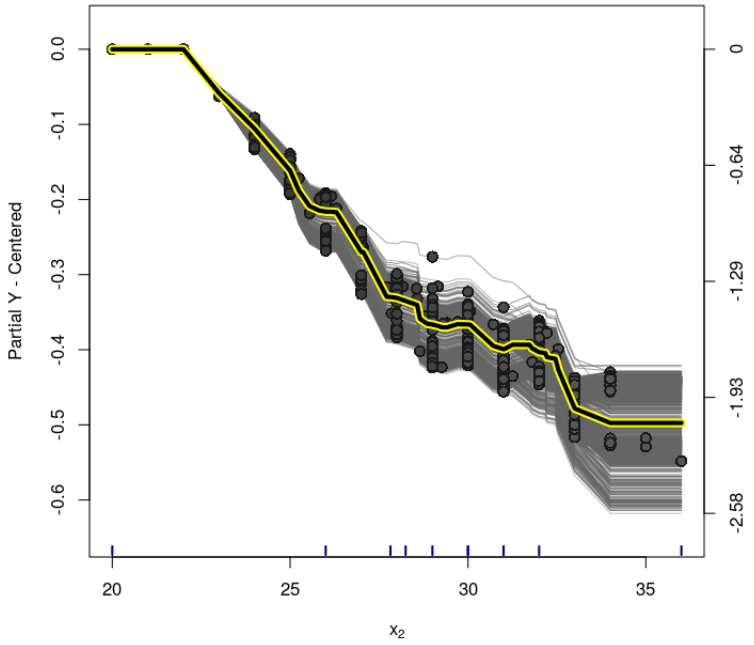

Figure C.131: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_2 - centered-ICE Plot for the treatment effect.*

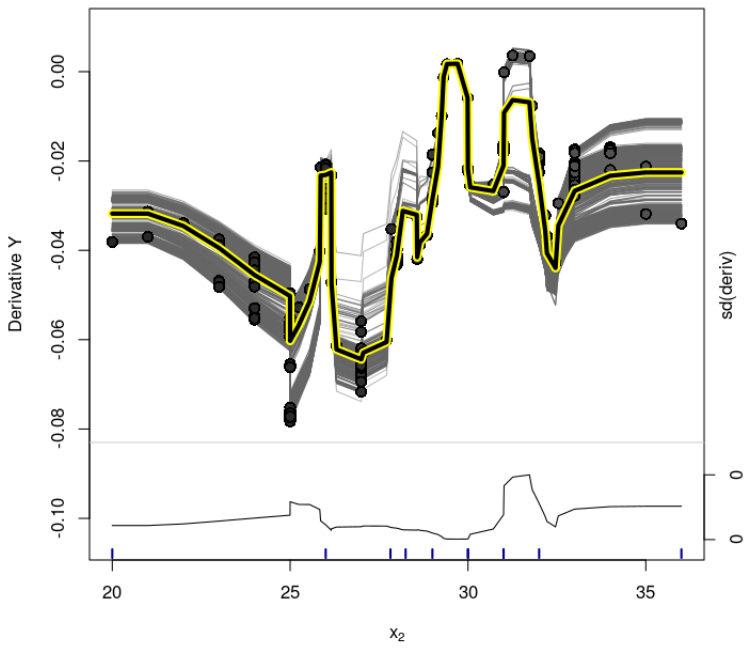

Figure C.132: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

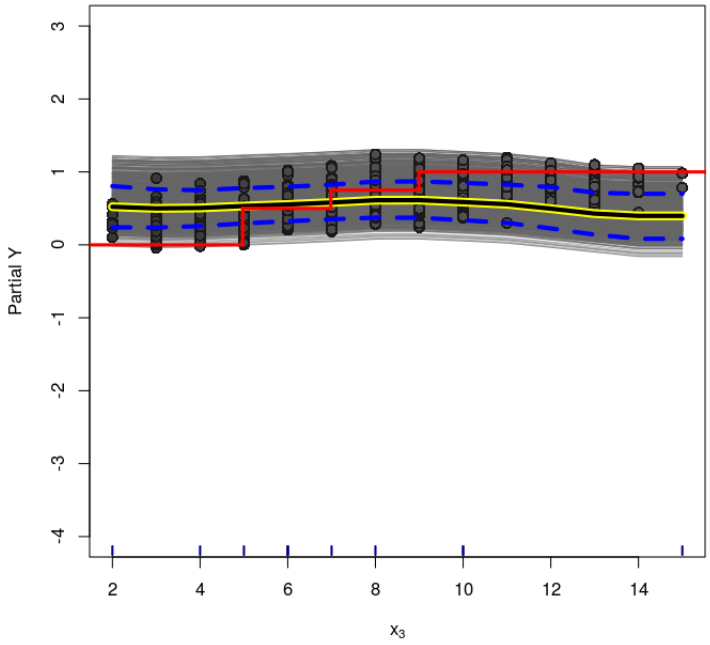

Figure C.133: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

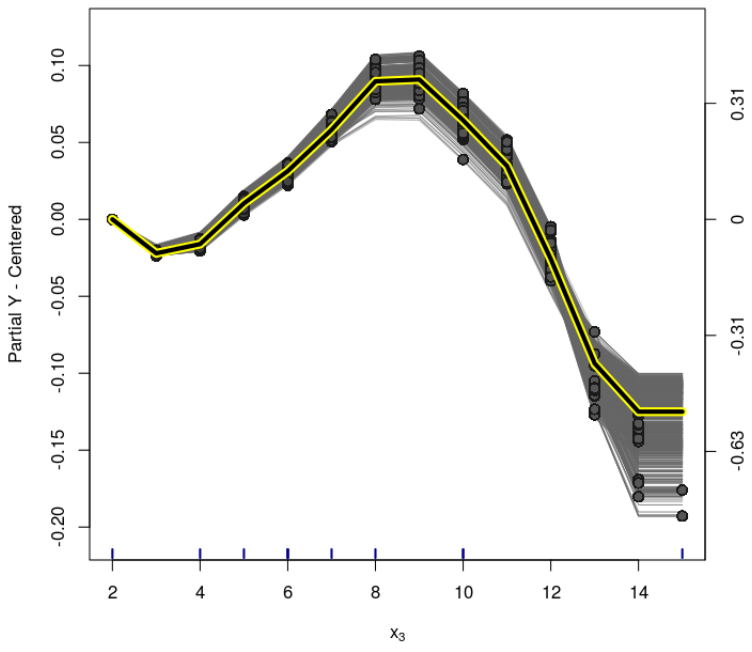

Figure C.134: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_3 - centered-ICE Plot for the treatment effect.*

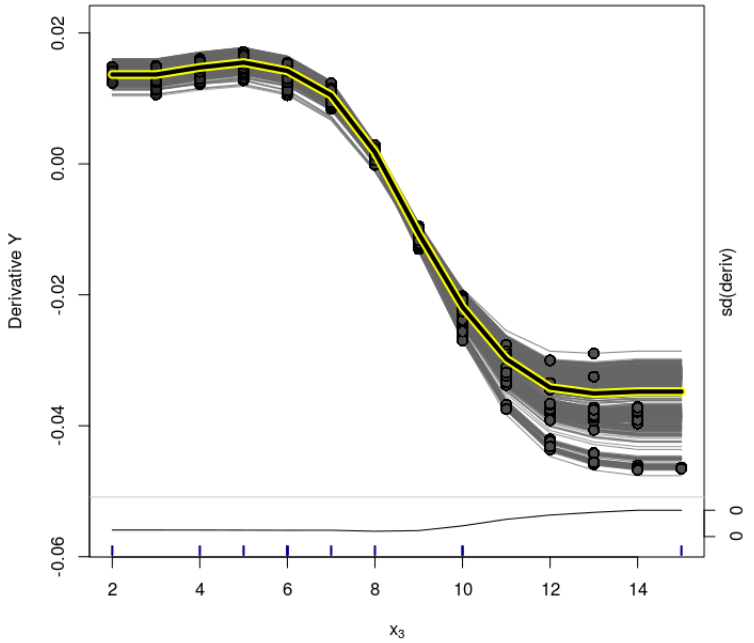

Figure C.135: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

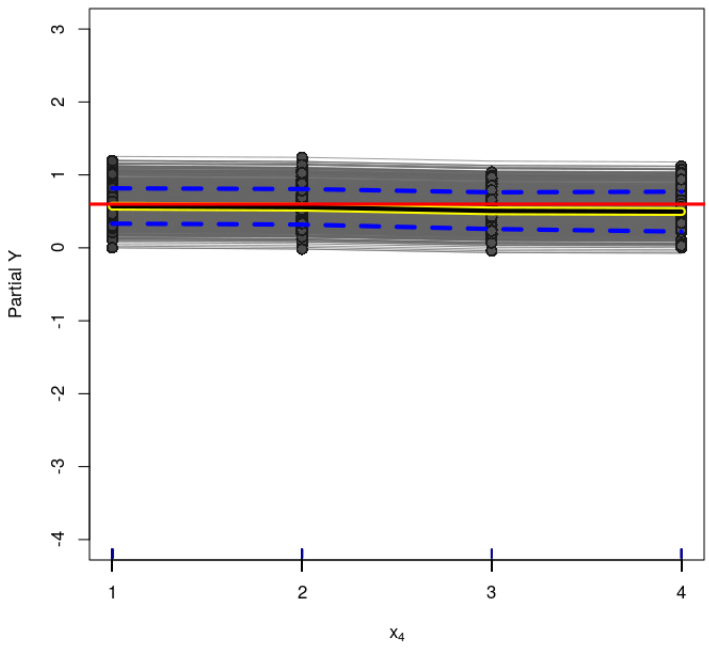

Figure C.136: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

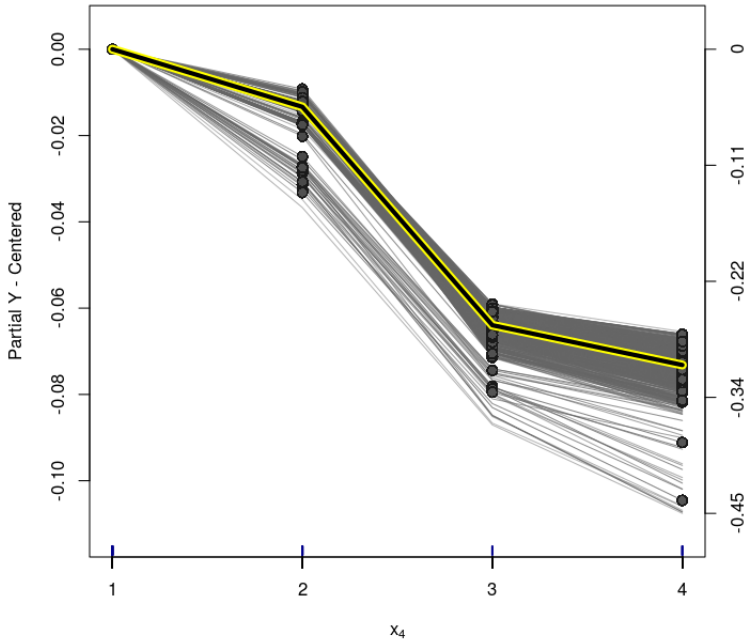

Figure C.137: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_4 - centered-ICE Plot for the treatment effect.*

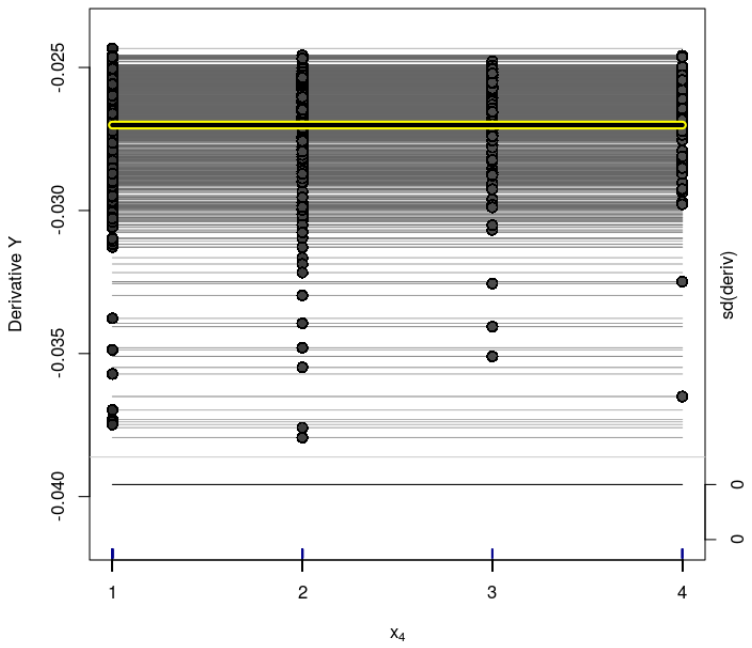

Figure C.138: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

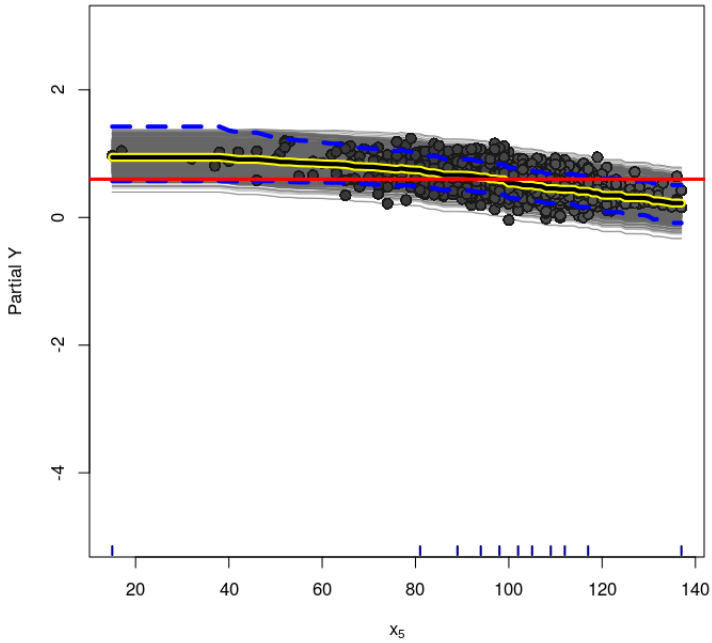

Figure C.139: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

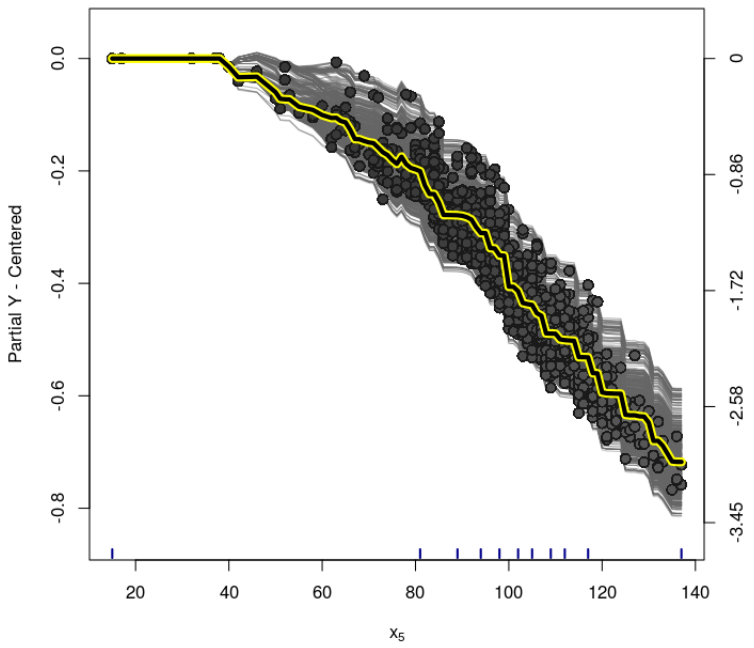

Figure C.140: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_5 - centered-ICE Plot for the treatment effect.*

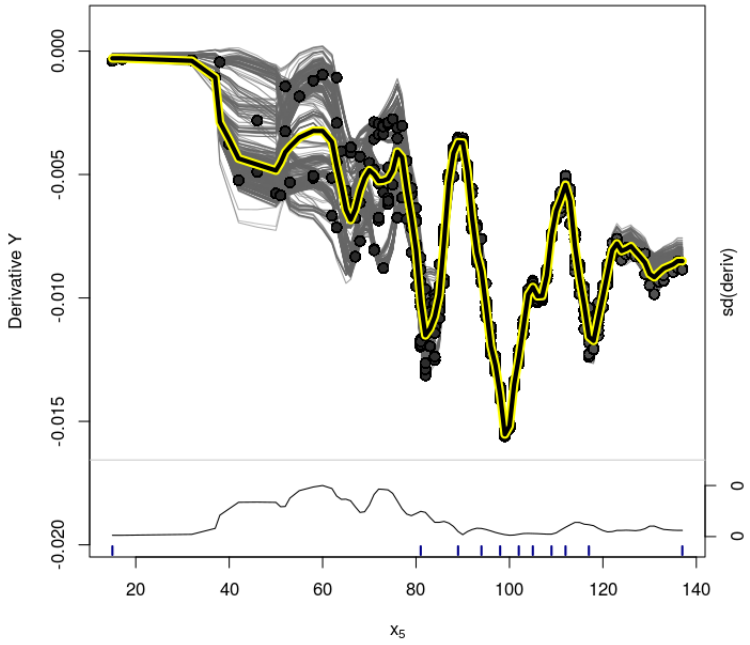

Figure C.141: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

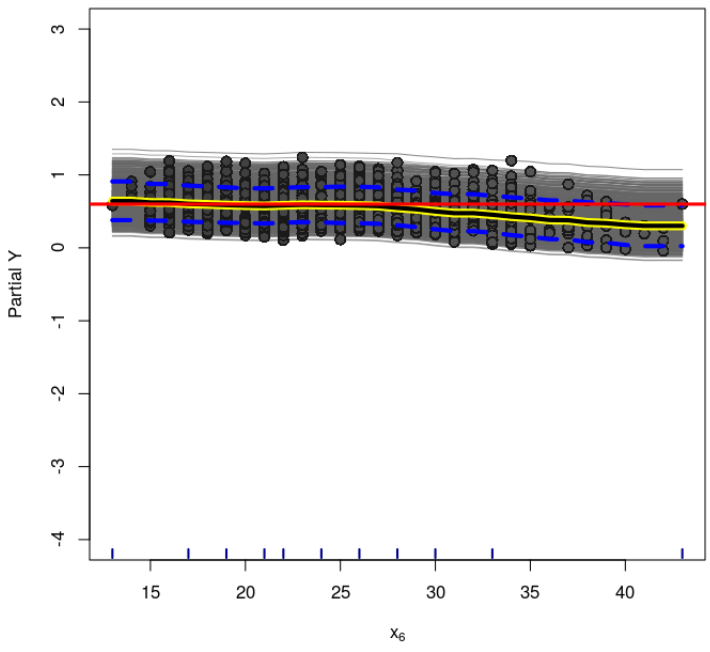

Figure C.142: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_6 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

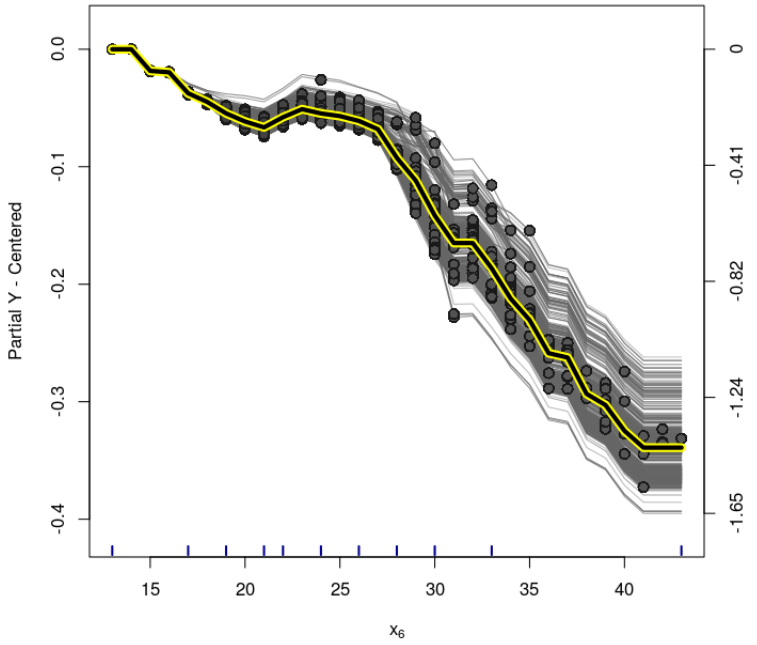

Figure C.143: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_6 - centered-ICE Plot for the treatment effect.*

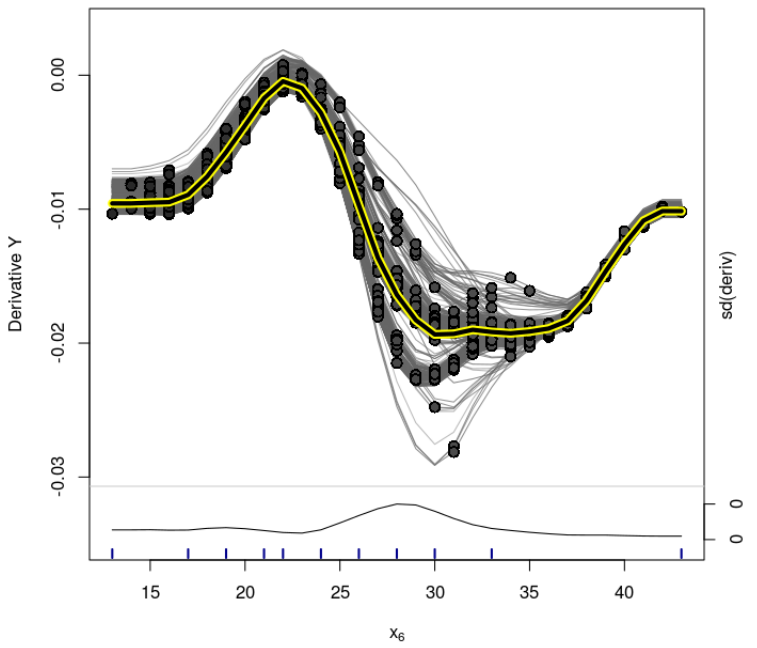

Figure C.144: *Oracle-BCF model (Simulation Based on Real Data) - Variable x_6 - d-ICE Plot for the treatment effect estimates.*

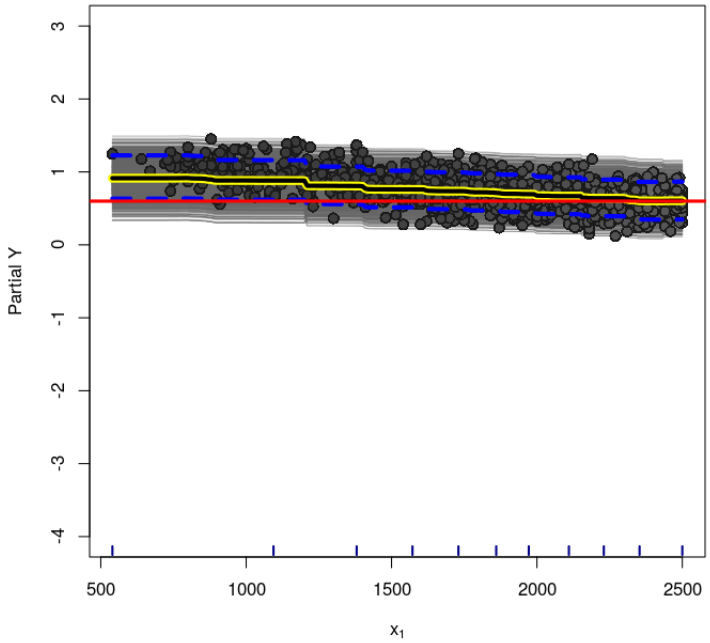

Figure C.145: *GLM-BCF model (Simulation Based on Real Data) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

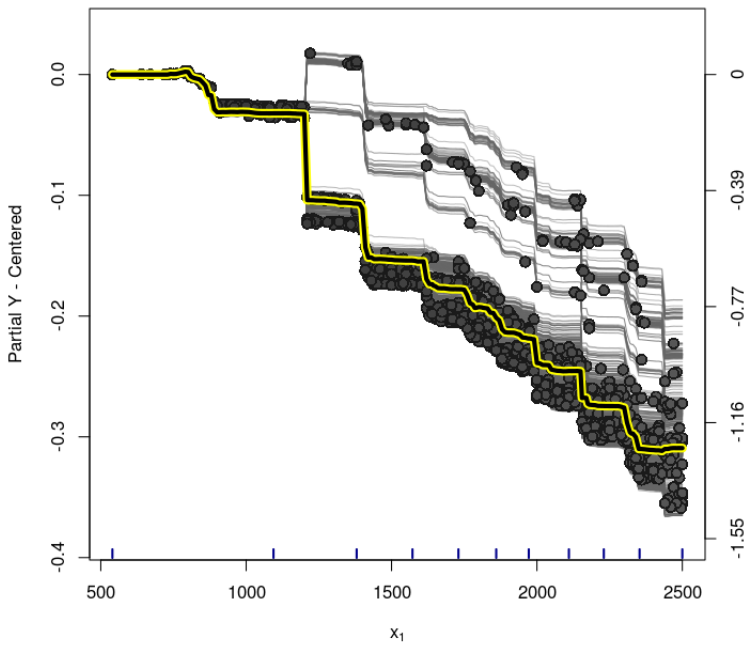

Figure C.146: *GLM-BCF model (Simulation Based on Real Data) - Variable x_1 - centered-ICE Plot for the treatment effect.*

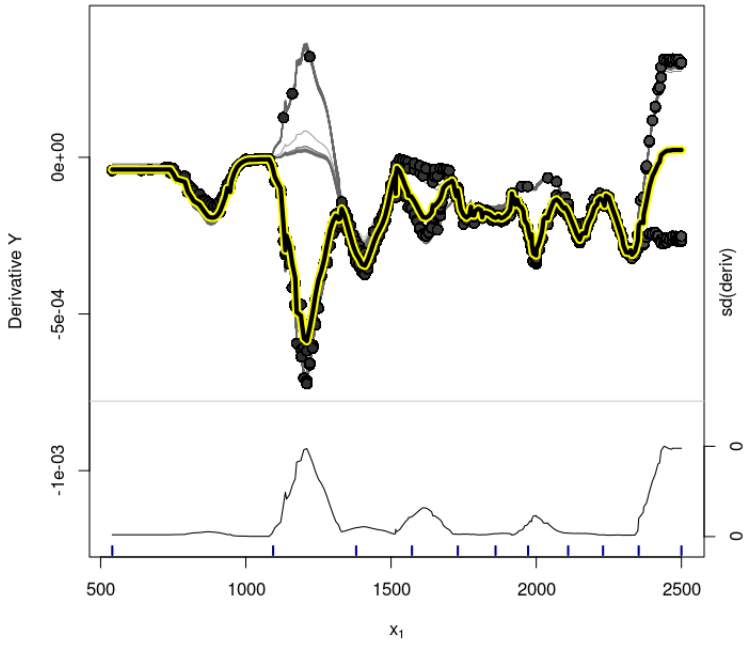

Figure C.147: *GLM-BCF model (Simulation Based on Real Data) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

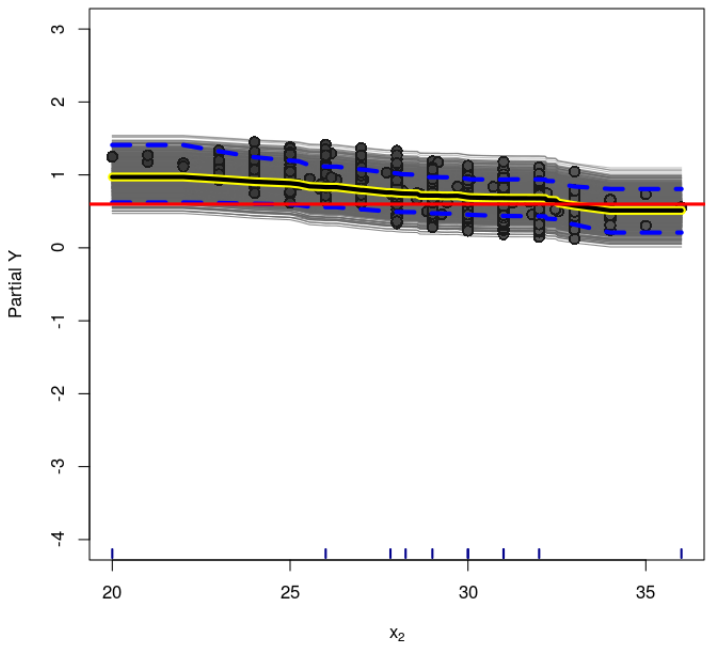

Figure C.148: *GLM-BCF model (Simulation Based on Real Data) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

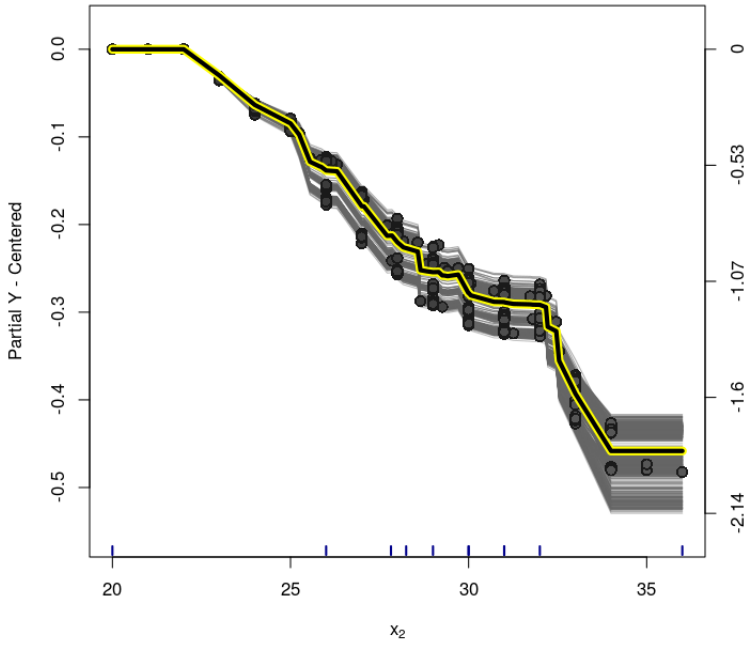

Figure C.149: *GLM-BCF model (Simulation Based on Real Data) - Variable x_2 - centered-ICE Plot for the treatment effect.*

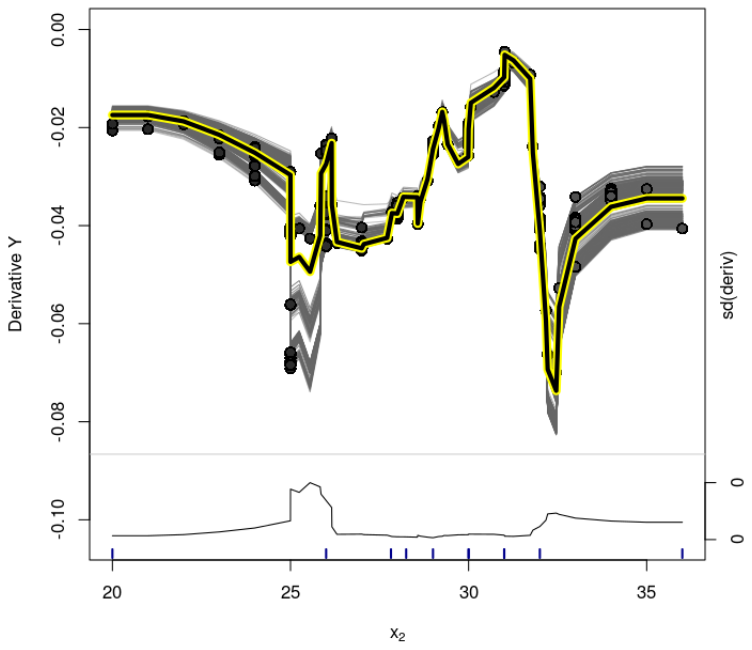

Figure C.150: *GLM-BCF model (Simulation Based on Real Data) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

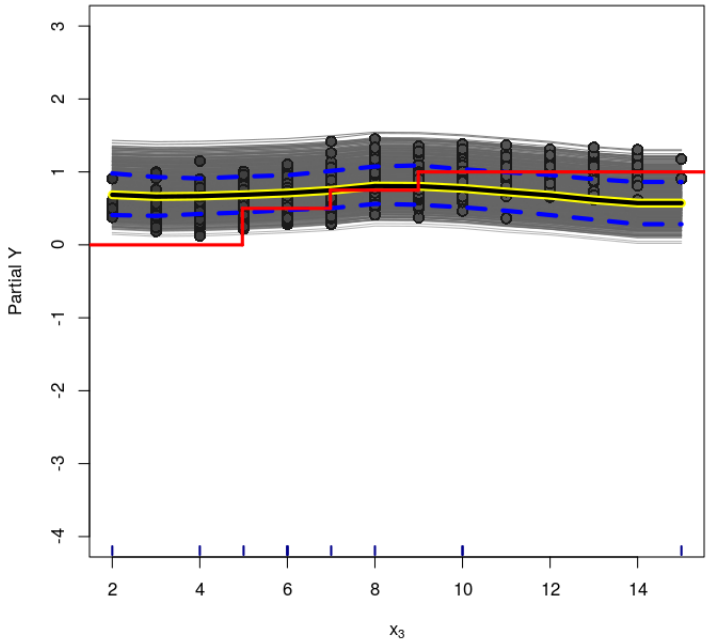

Figure C.151: *GLM-BCF model (Simulation Based on Real Data) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

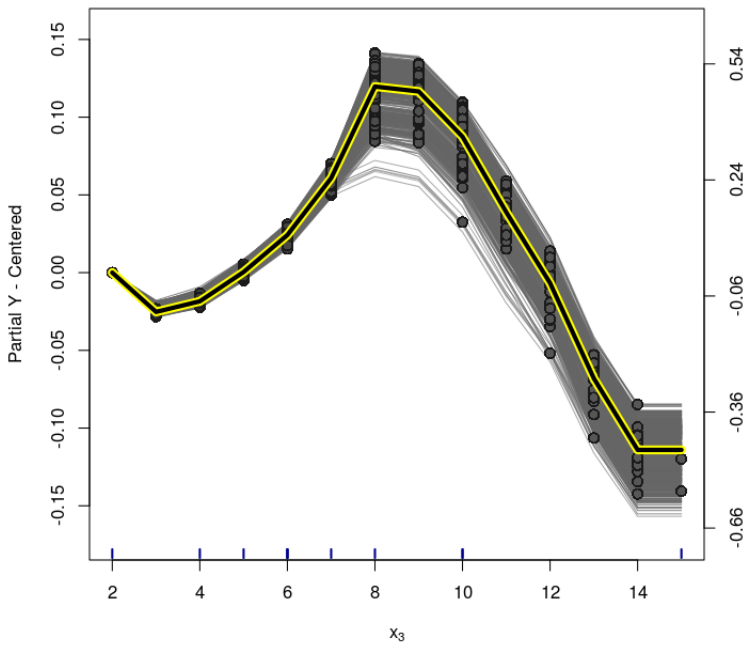

Figure C.152: *GLM-BCF model (Simulation Based on Real Data) - Variable x_3 - centered-ICE Plot for the treatment effect.*

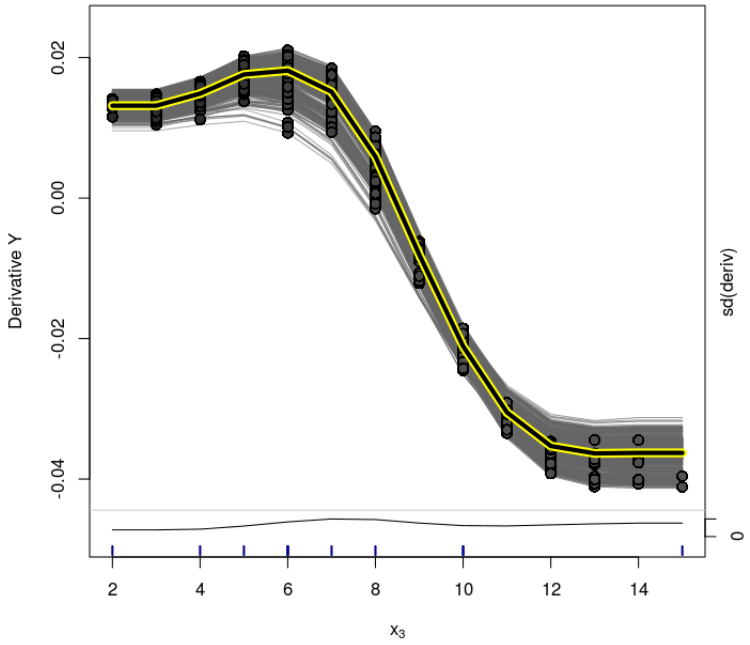

Figure C.153: *GLM-BCF model (Simulation Based on Real Data) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

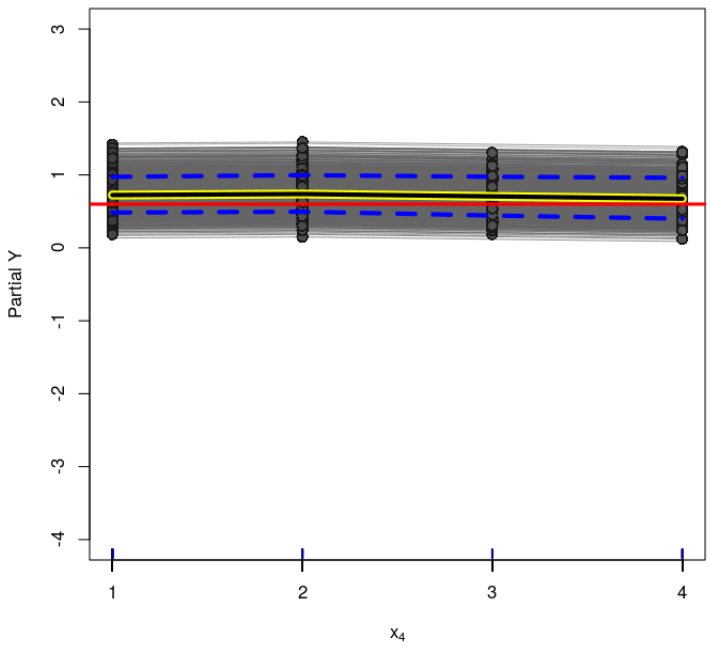

Figure C.154: *GLM-BCF model (Simulation Based on Real Data) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

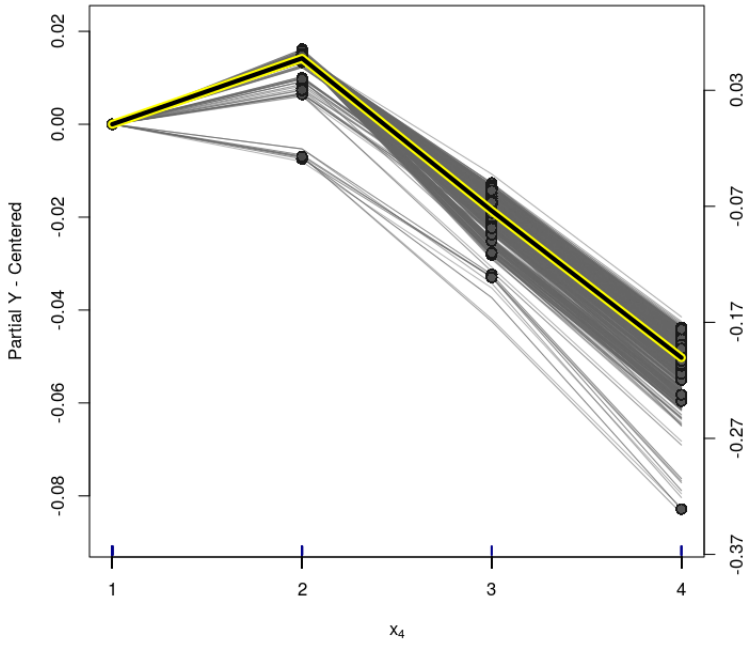

Figure C.155: *GLM-BCF model (Simulation Based on Real Data) - Variable x_4 - centered-ICE Plot for the treatment effect.*

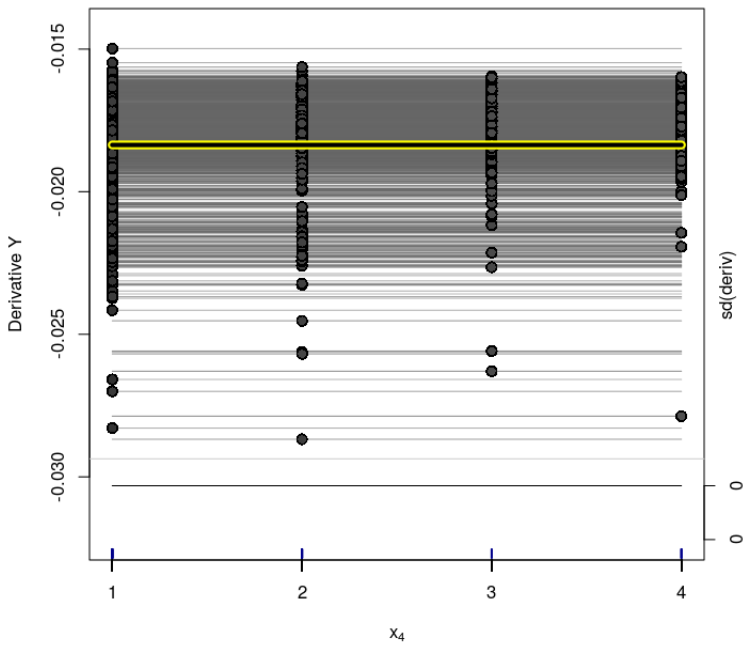

Figure C.156: *GLM-BCF model (Simulation Based on Real Data) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

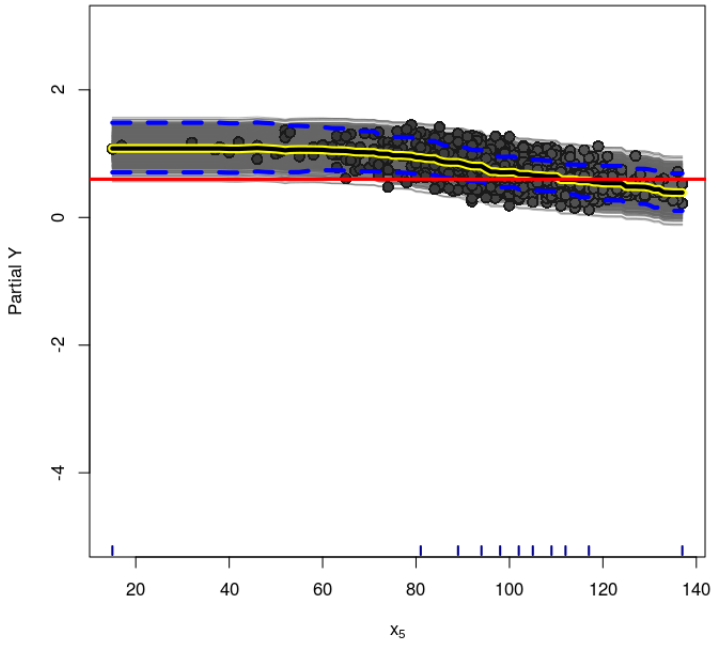

Figure C.157: *GLM-BCF model (Simulation Based on Real Data) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

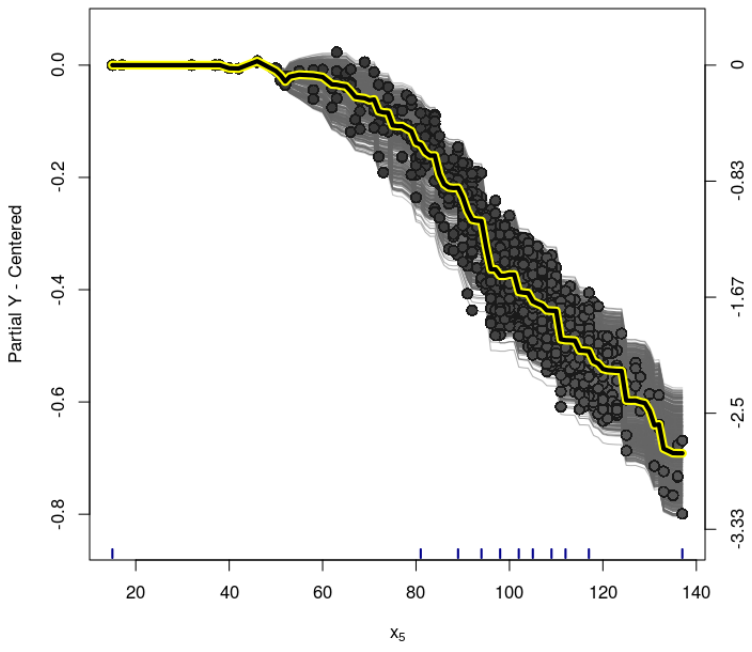

Figure C.158: *GLM-BCF model (Simulation Based on Real Data) - Variable x_5 - centered-ICE Plot for the treatment effect.*

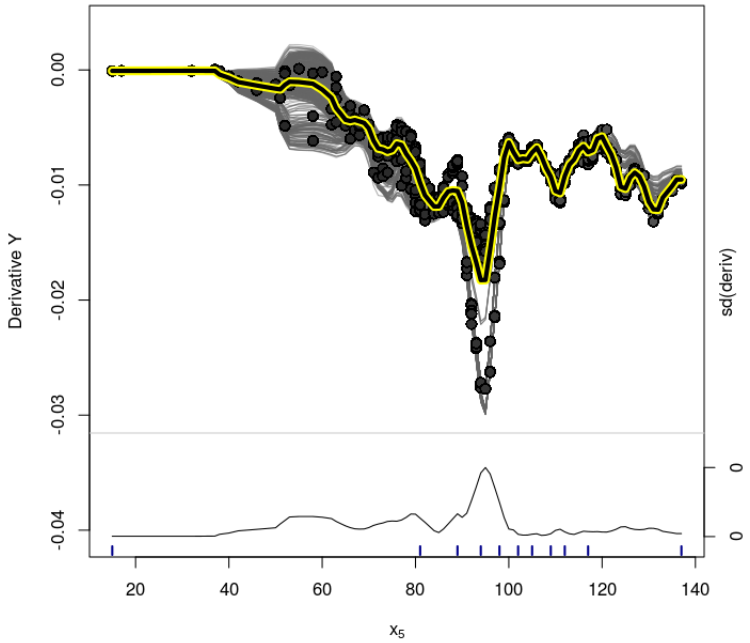

Figure C.159: *GLM-BCF model (Simulation Based on Real Data) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

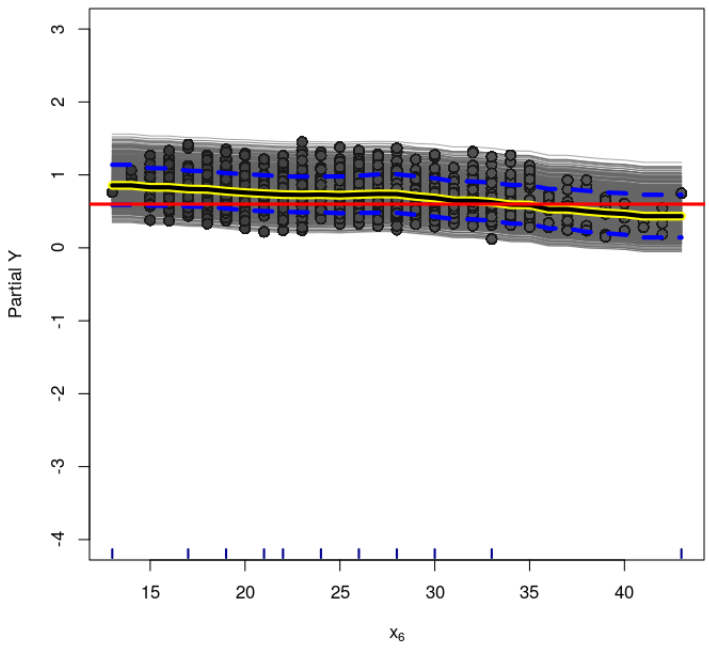

Figure C.160: *GLM-BCF model (Simulation Based on Real Data) - Variable x_6 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

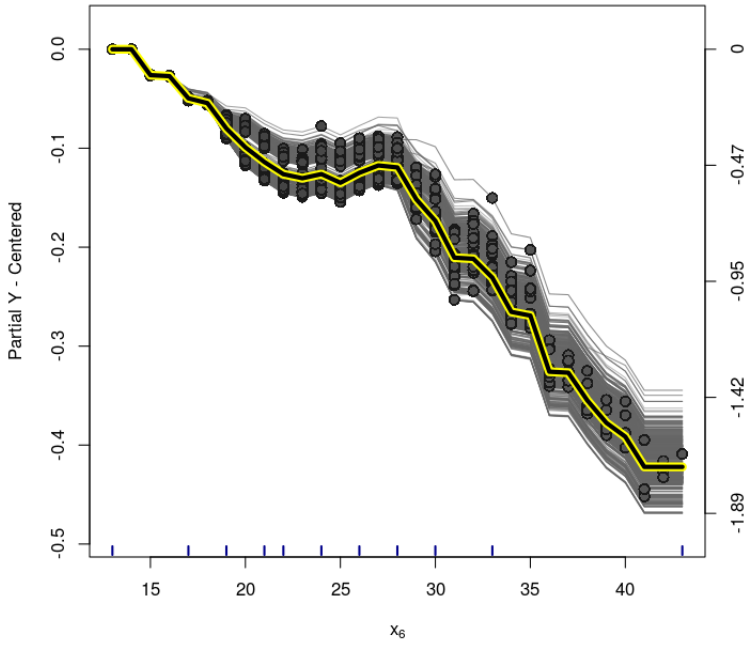

Figure C.161: *GLM-BCF model (Simulation Based on Real Data) - Variable x_6 - centered-ICE Plot for the treatment effect.*

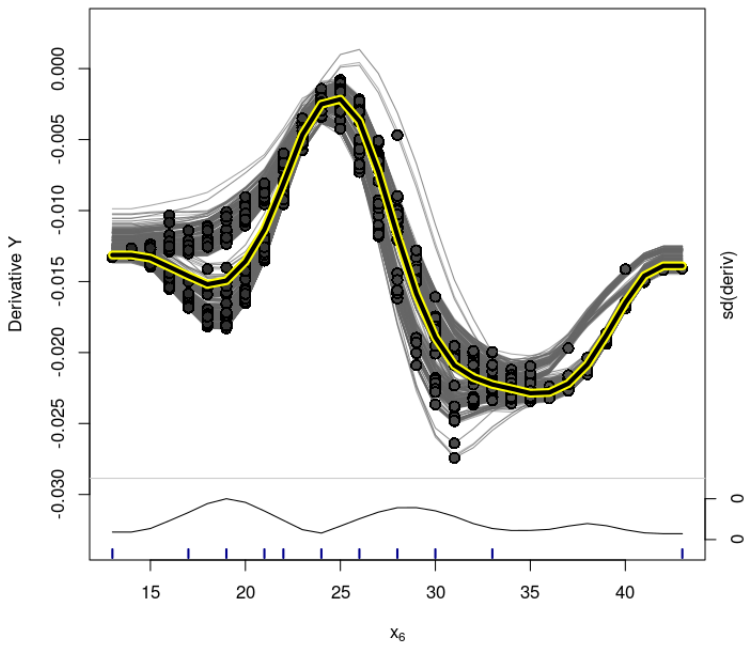

Figure C.162: *GLM-BCF model (Simulation Based on Real Data) - Variable x_6 - d-ICE Plot for the treatment effect estimates.*

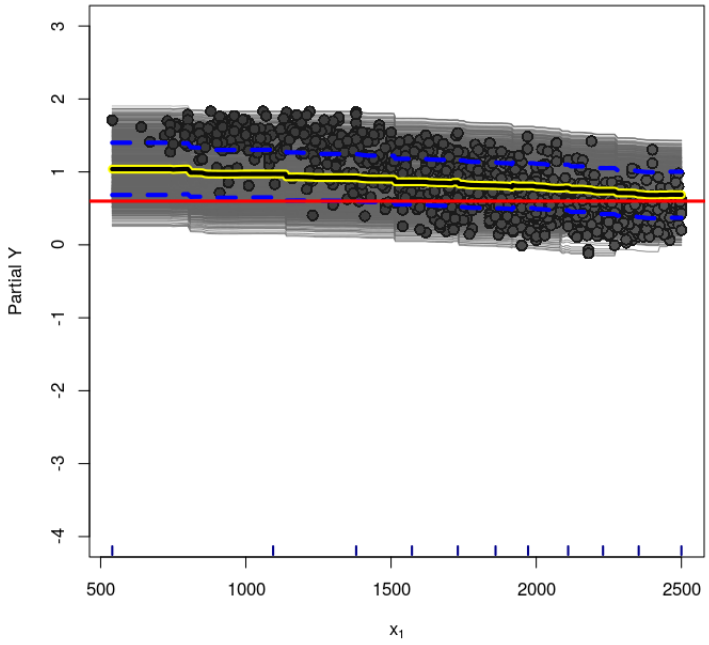

Figure C.163: *Rand-BCF model (Simulation Based on Real Data) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

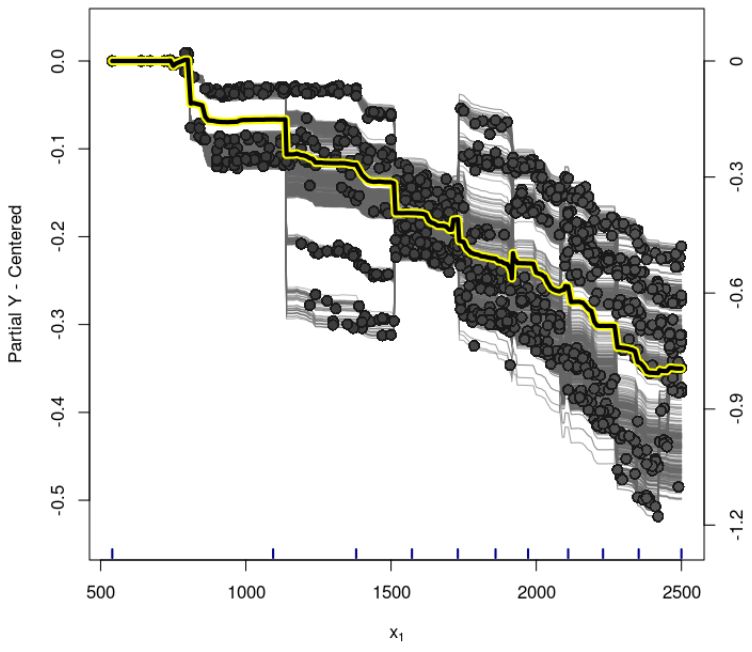

Figure C.164: *Rand-BCF model (Simulation Based on Real Data) - Variable x_1 - centered-ICE Plot for the treatment effect.*

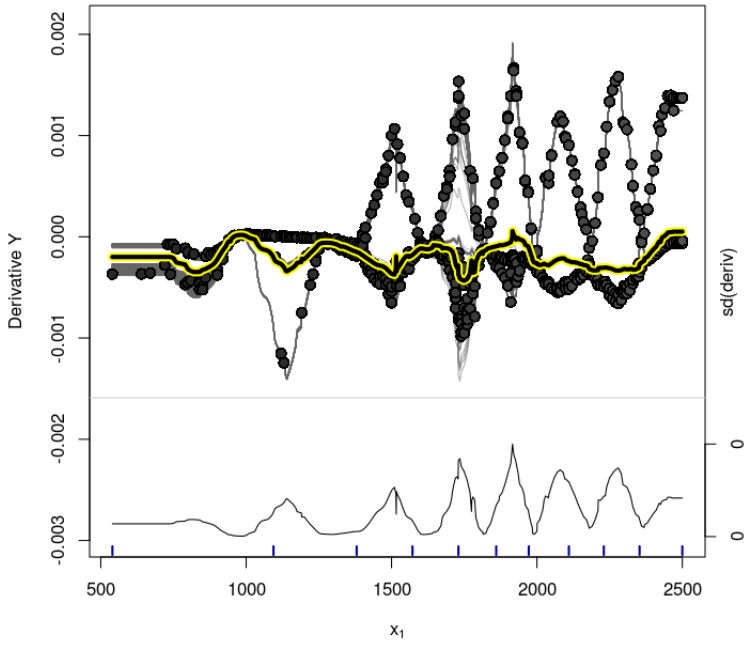

Figure C.165: *Rand-BCF model (Simulation Based on Real Data) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

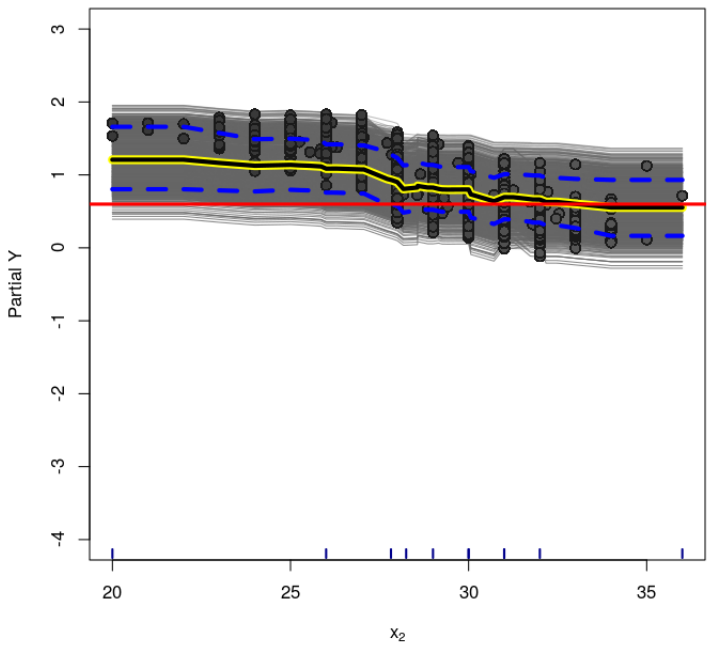

Figure C.166: *Rand-BCF model (Simulation Based on Real Data) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

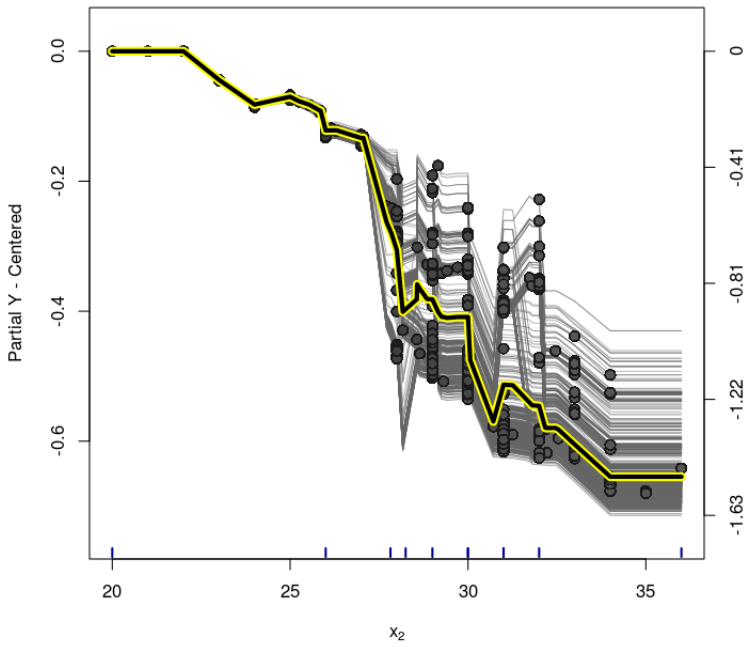

Figure C.167: *Rand-BCF model (Simulation Based on Real Data) - Variable x_2 - centered-ICE Plot for the treatment effect.*

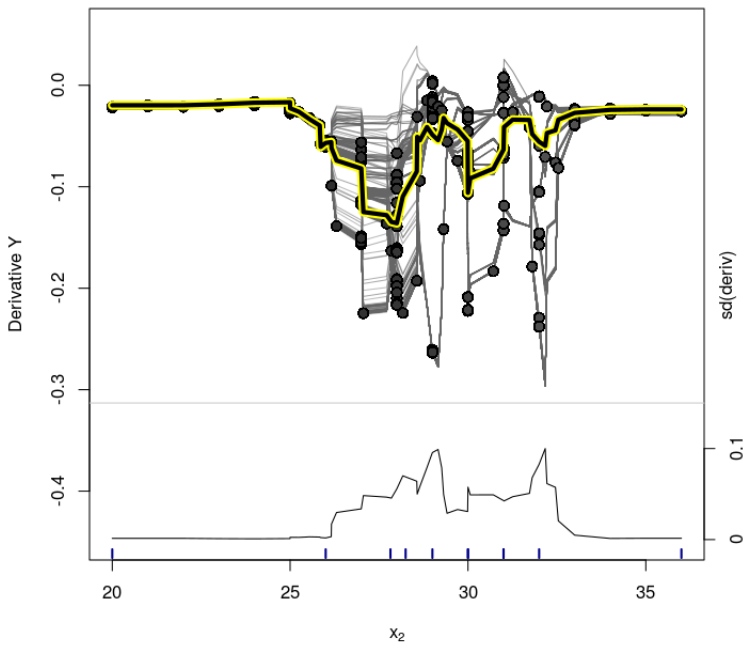

Figure C.168: *Rand-BCF model (Simulation Based on Real Data) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

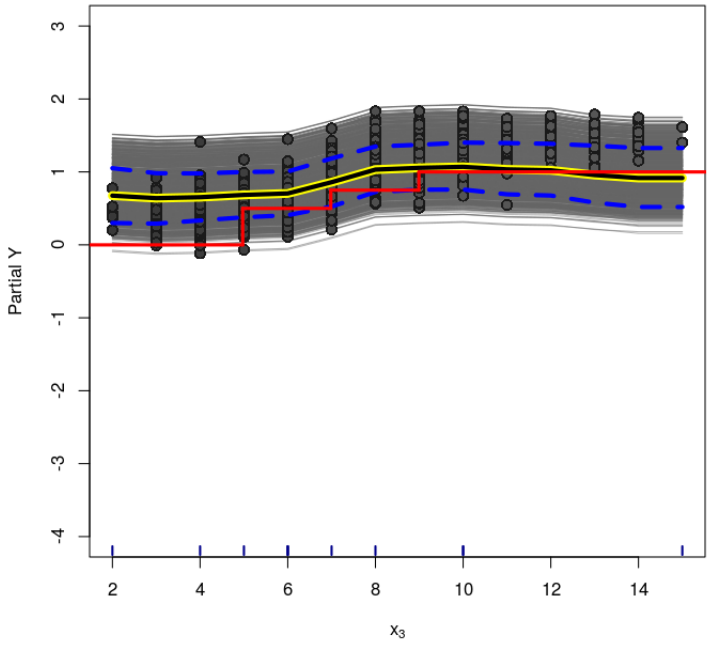

Figure C.169: *Rand-BCF model (Simulation Based on Real Data) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

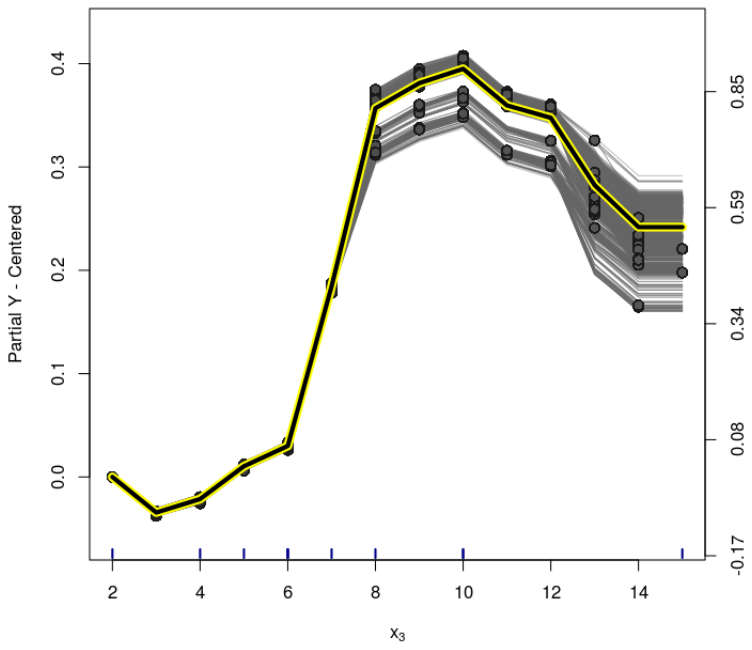

Figure C.170: *Rand-BCF model (Simulation Based on Real Data) - Variable x_3 - centered-ICE Plot for the treatment effect.*

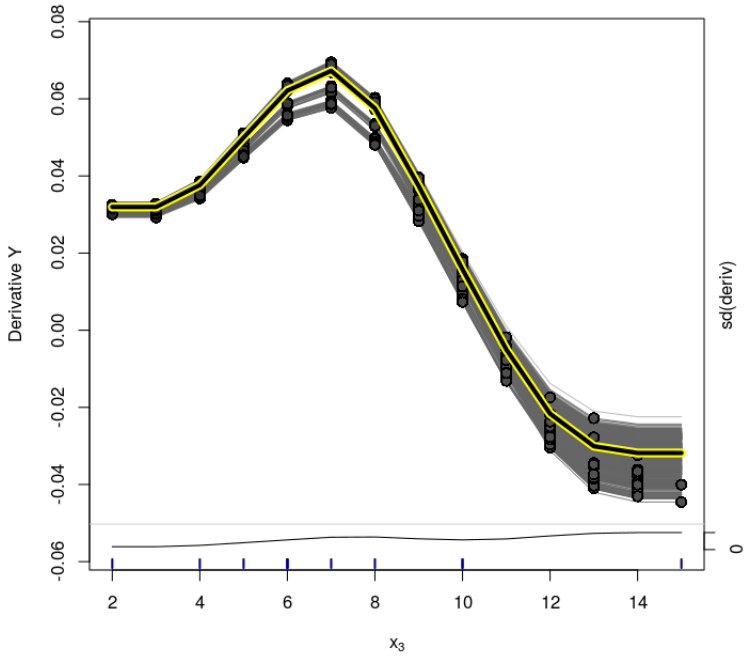

Figure C.171: *Rand-BCF model (Simulation Based on Real Data) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

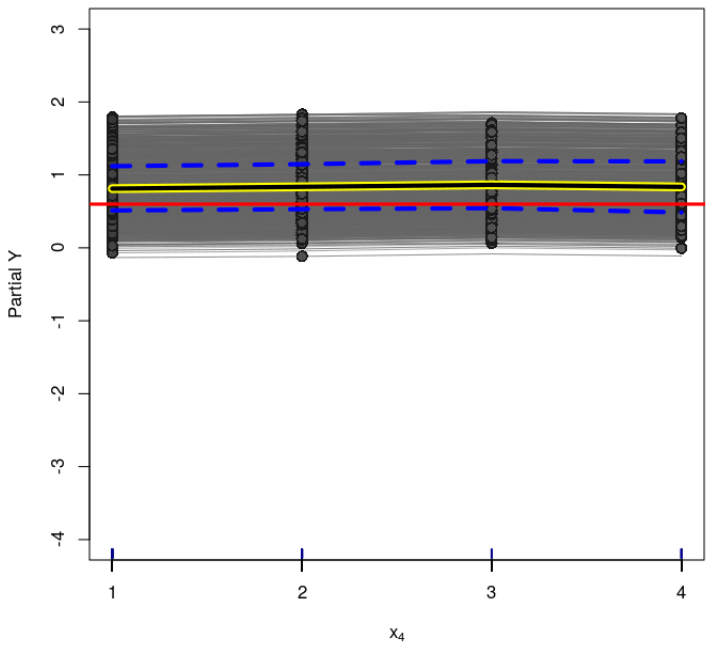

Figure C.172: *Rand-BCF model (Simulation Based on Real Data) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

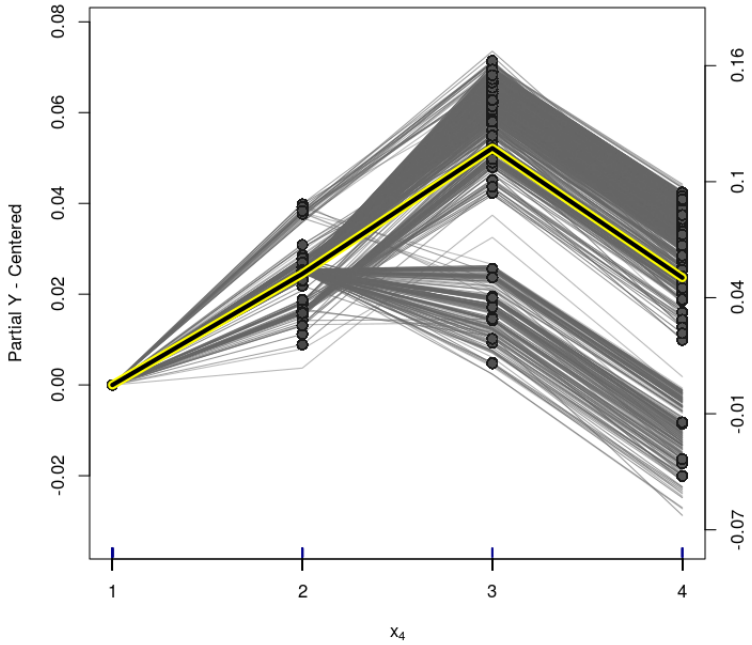

Figure C.173: *Rand-BCF model (Simulation Based on Real Data) - Variable x_4 - centered-ICE Plot for the treatment effect.*

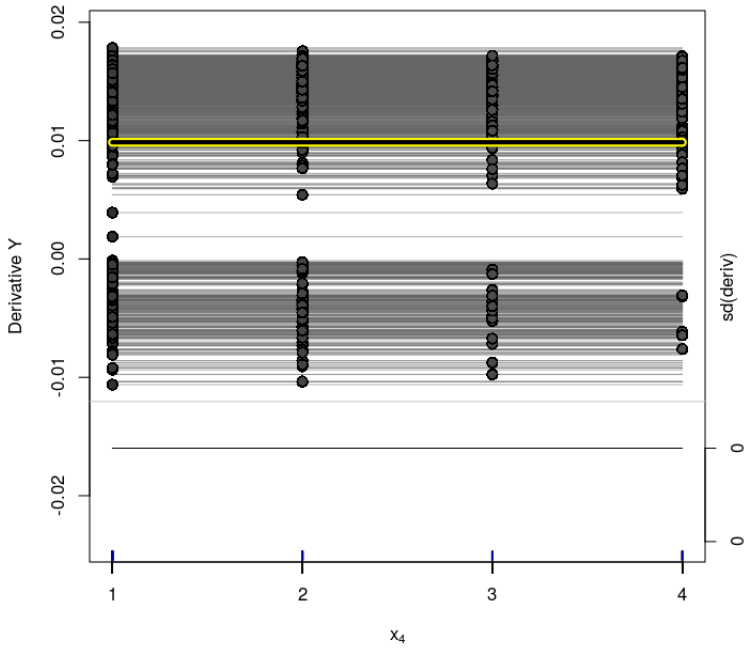

Figure C.174: *Rand-BCF model (Simulation Based on Real Data) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

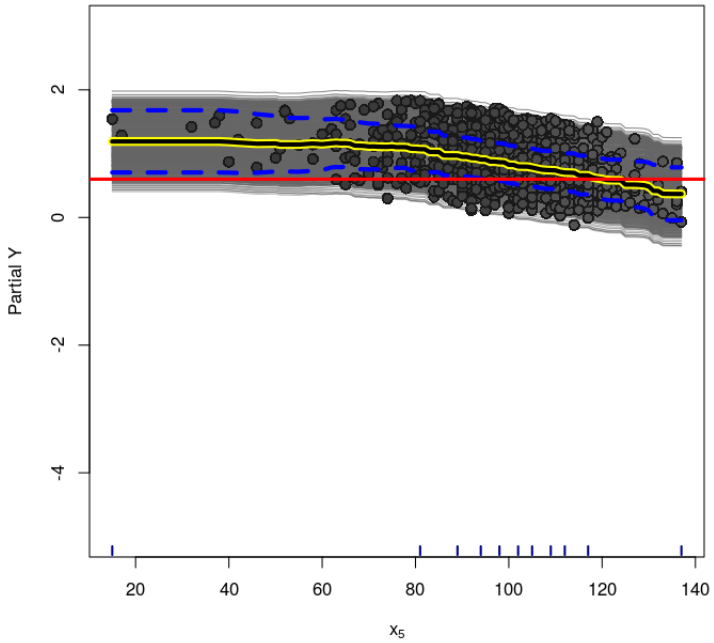

Figure C.175: *Rand-BCF model (Simulation Based on Real Data) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

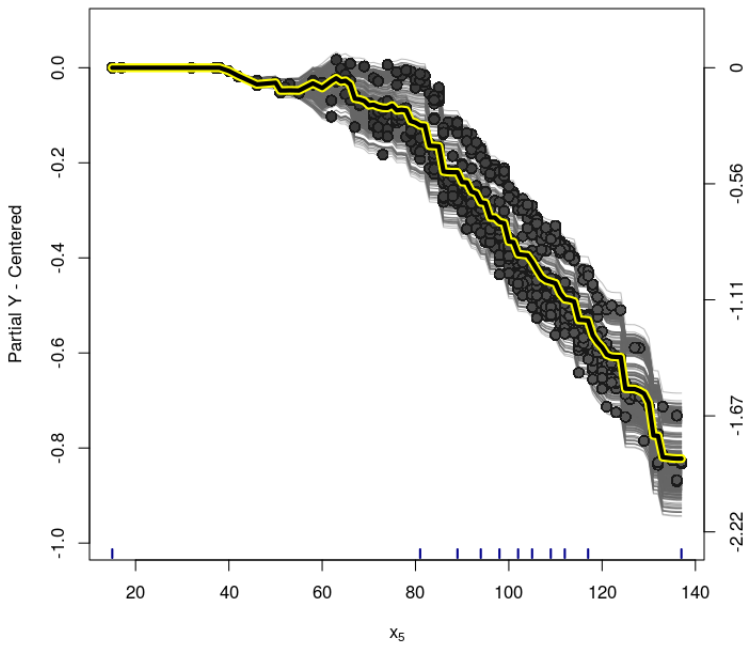

Figure C.176: *Rand-BCF model (Simulation Based on Real Data) - Variable x_5 - centered-ICE Plot for the treatment effect.*

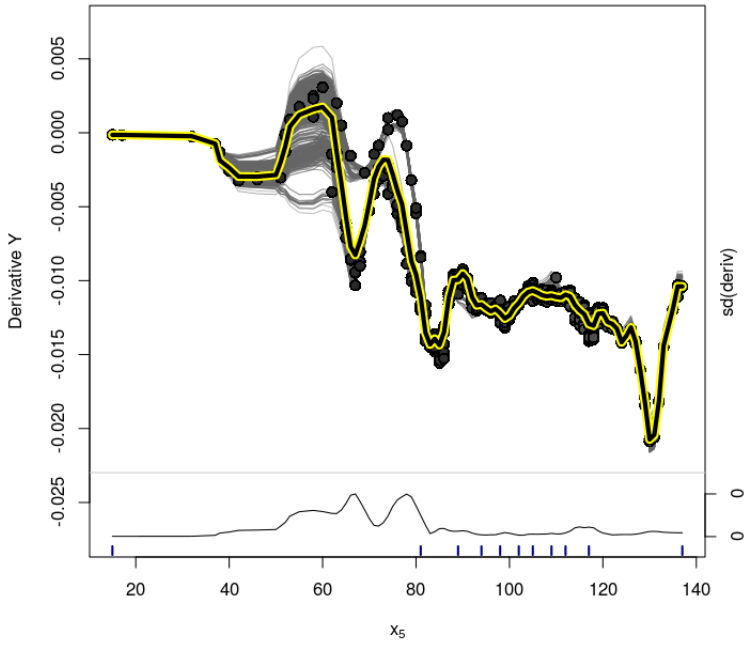

Figure C.177: *Rand-BCF model (Simulation Based on Real Data) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

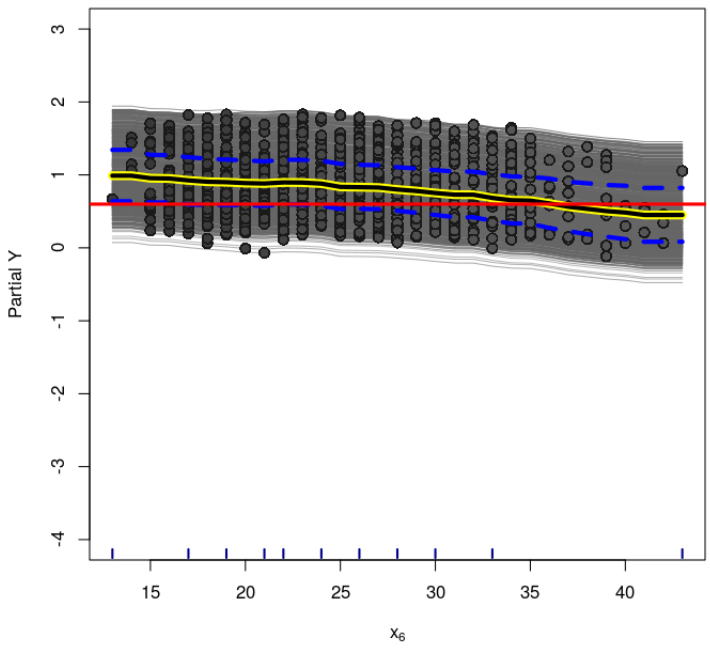

Figure C.178: *Rand-BCF model (Simulation Based on Real Data) - Variable x_6 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

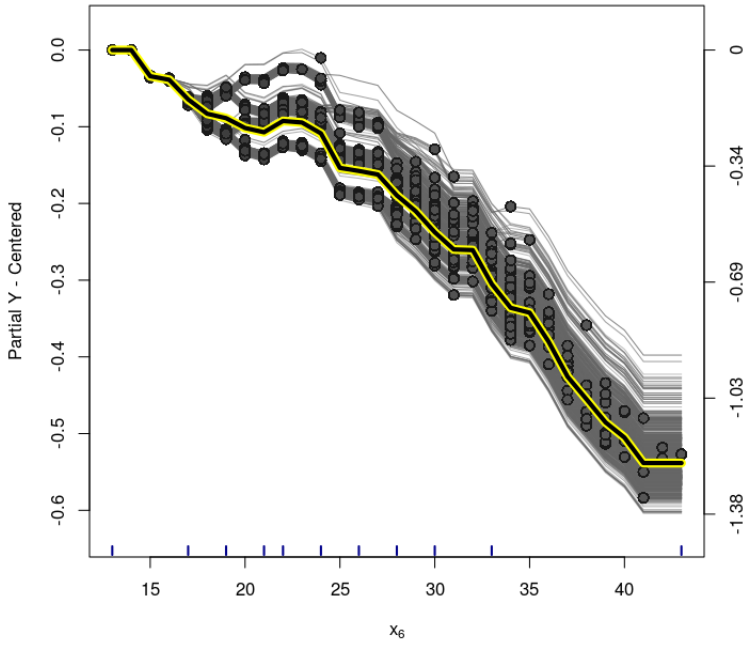

Figure C.179: *Rand-BCF model (Simulation Based on Real Data) - Variable x_6 - centered-ICE Plot for the treatment effect.*

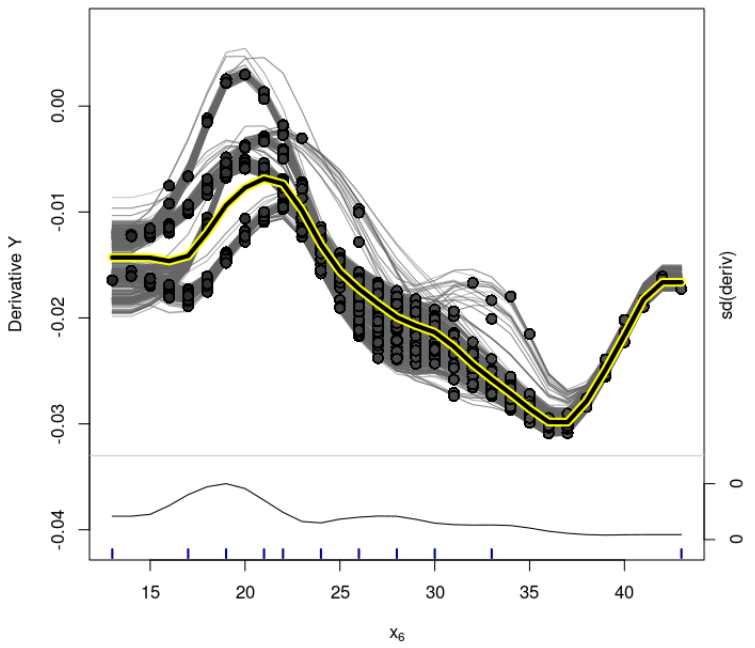

Figure C.180: *Rand-BCF model (Simulation Based on Real Data) - Variable x_6 - d-ICE Plot for the treatment effect estimates.*

Appendix D

Graphics - Friedman Function under Sparsity

D.1 Convergence Analysis and Variable Selection

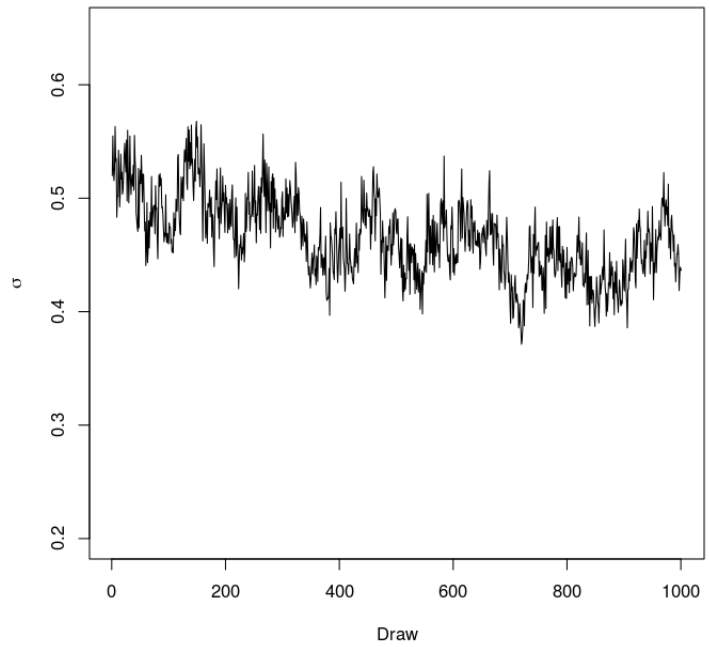

Figure D.1: *Vanilla model (Friedman Function under Sparsity) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

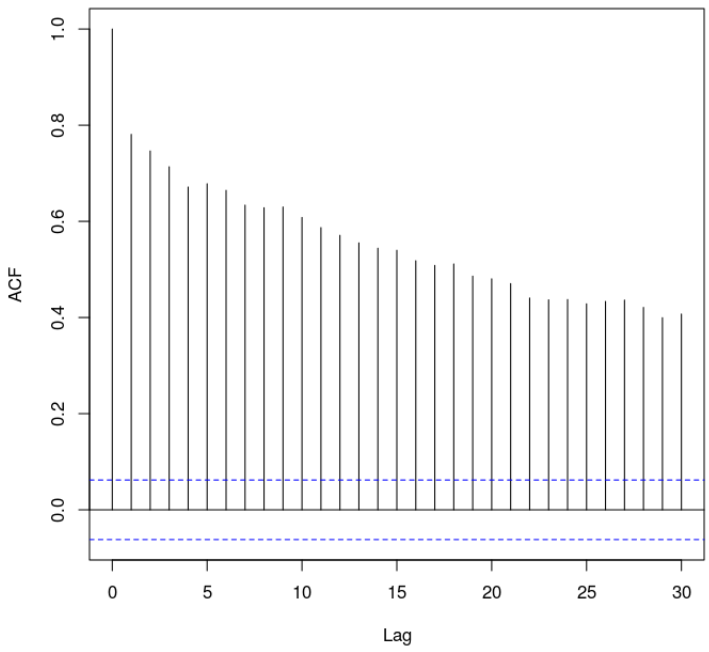

Figure D.2: *Vanilla model (Friedman Function under Sparsity) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

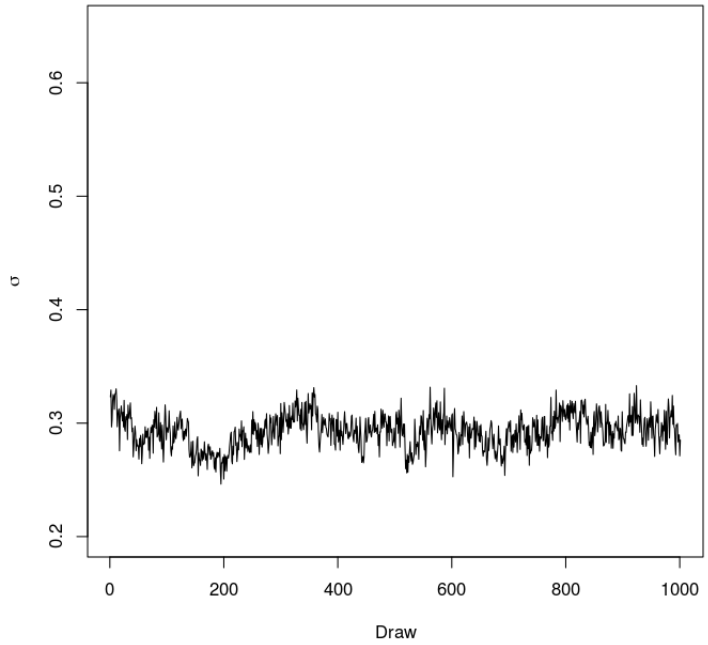

Figure D.3: *Oracle model (Friedman Function under Sparsity) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

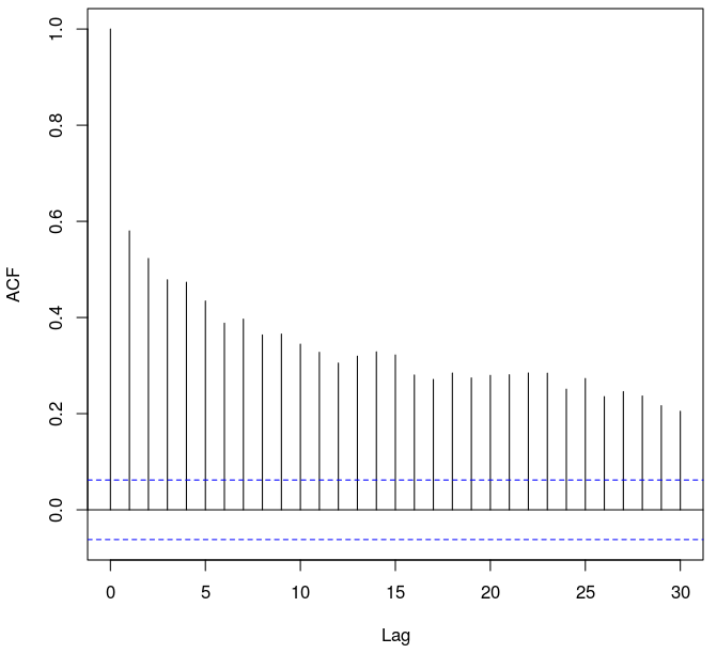

Figure D.4: *Oracle model (Friedman Function under Sparsity) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

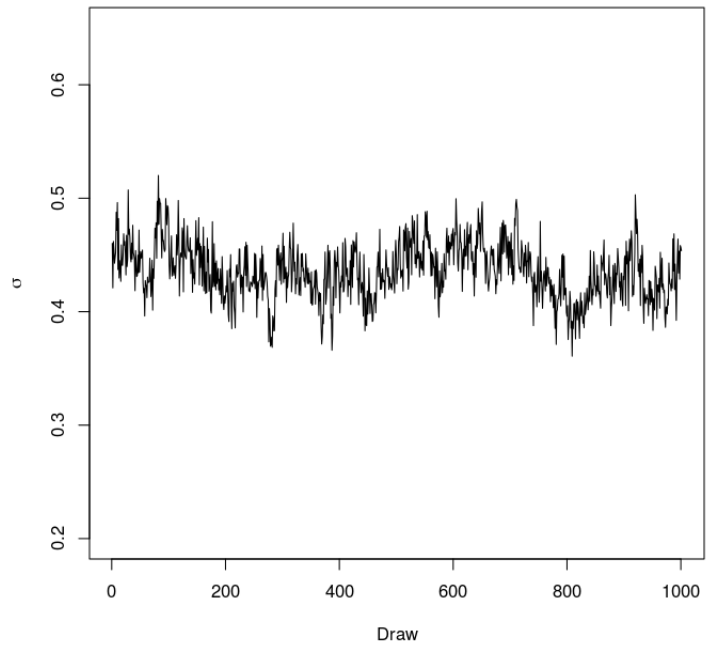

Figure D.5: *PS-BART model (Friedman Function under Sparsity) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

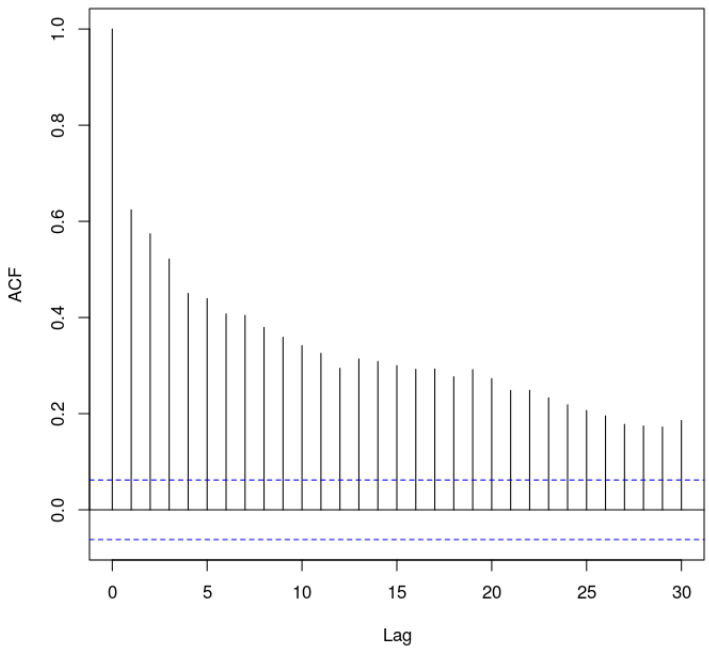

Figure D.6: *PS-BART model (Friedman Function under Sparsity) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

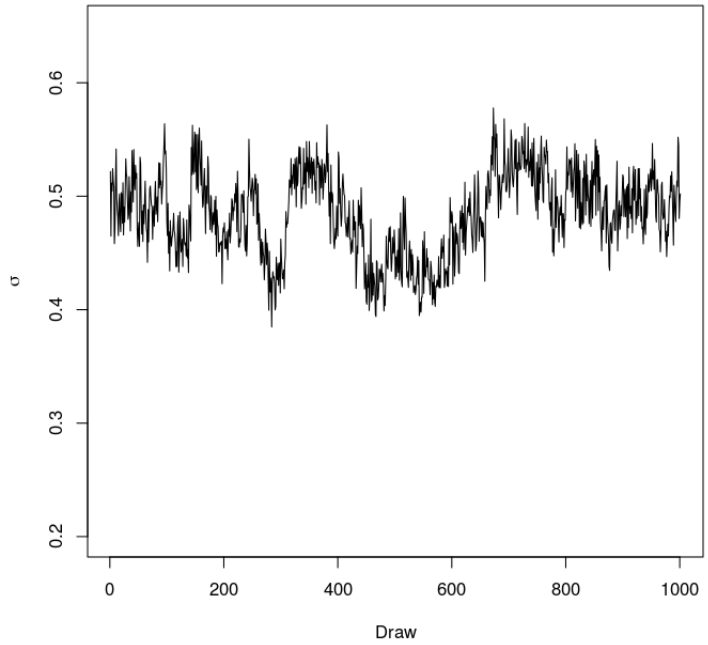

Figure D.7: *GLM-BART model (Friedman Function under Sparsity) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

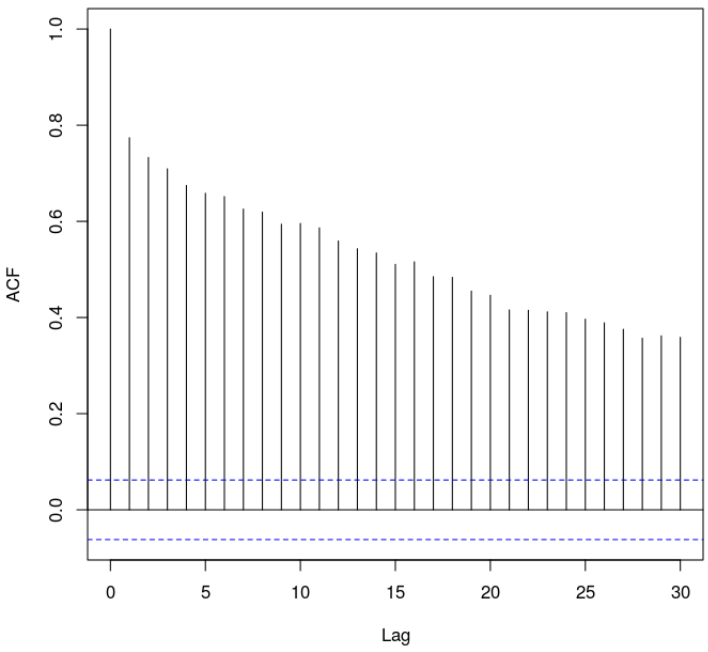

Figure D.8: *GLM-BART model (Friedman Function under Sparsity) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

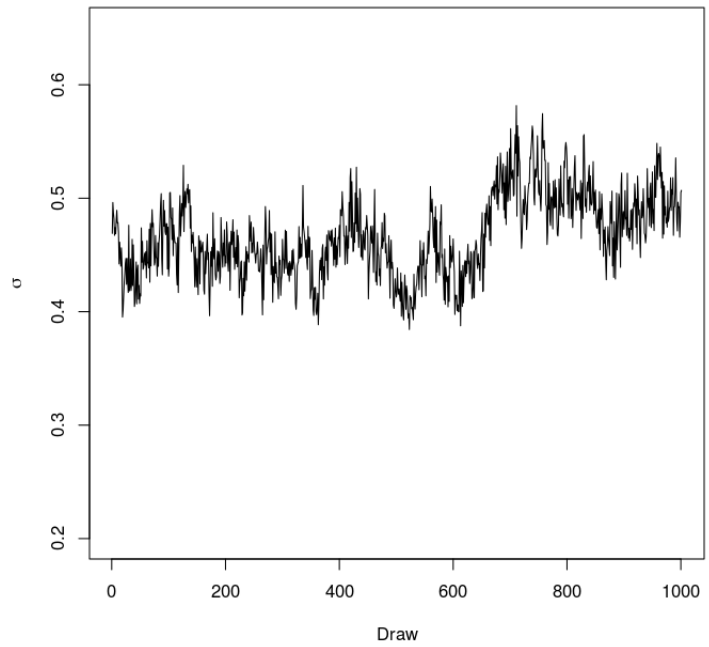

Figure D.9: *Rand-BART model (Friedman Function under Sparsity) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

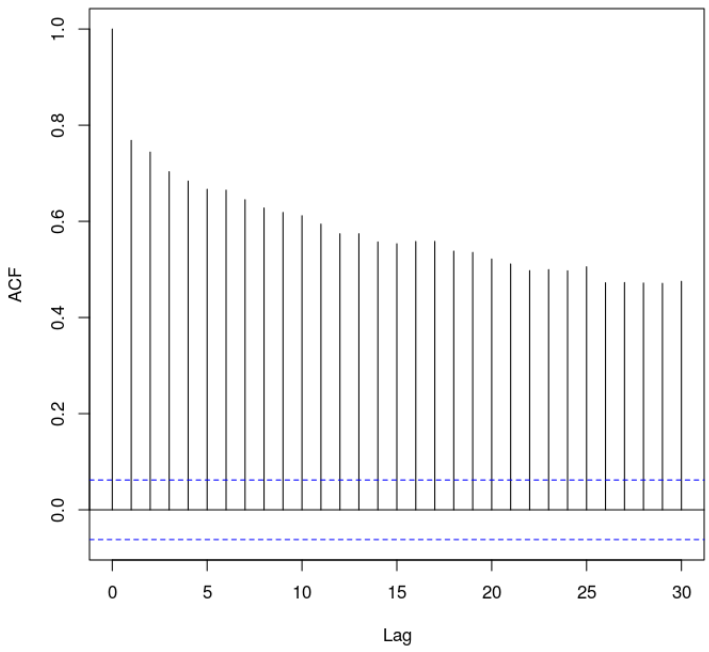

Figure D.10: *Rand-BART model (Friedman Function under Sparsity) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

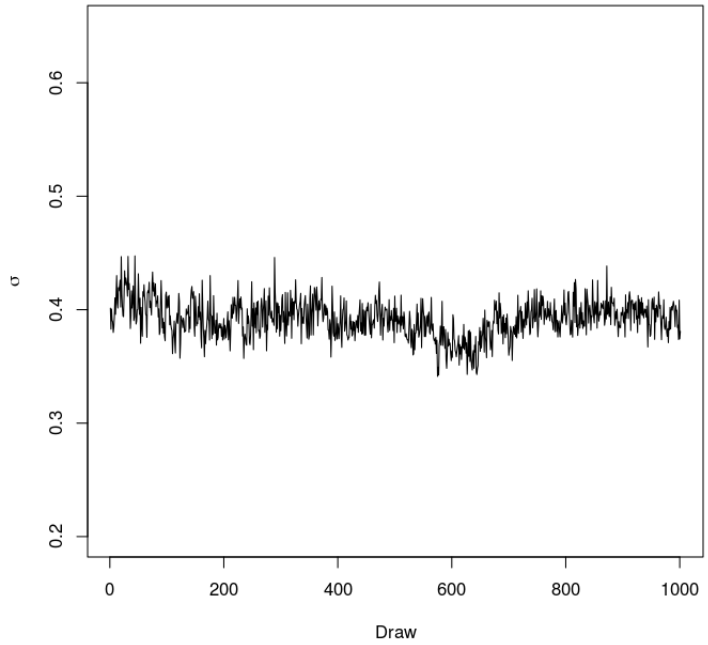

Figure D.11: *Vanilla-DART model (Friedman Function under Sparsity) - σ posterior draws trace plot. Apparently the draws traverse the sample space adequately.*

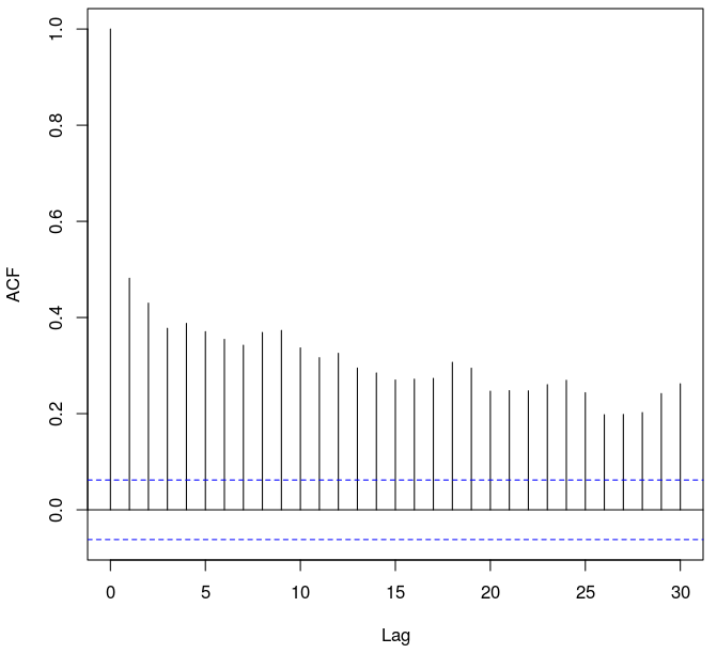

Figure D.12: *Vanilla-DART model (Friedman Function under Sparsity) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

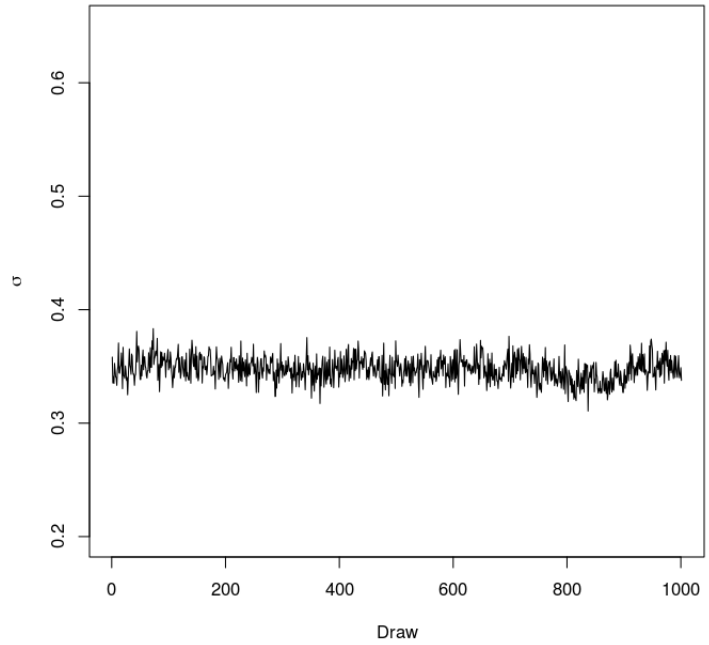

Figure D.13: *Oracle-DART model (Friedman Function under Sparsity) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

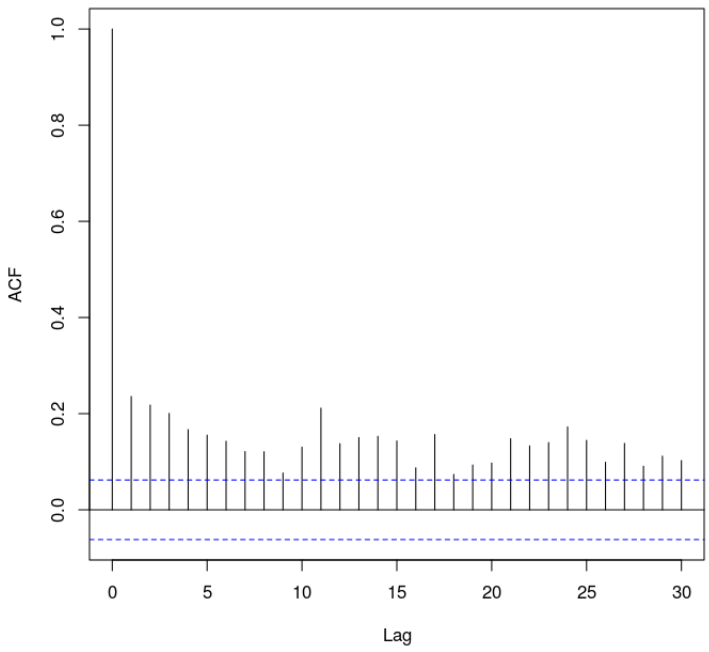

Figure D.14: *Oracle-DART model (Friedman Function under Sparsity) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

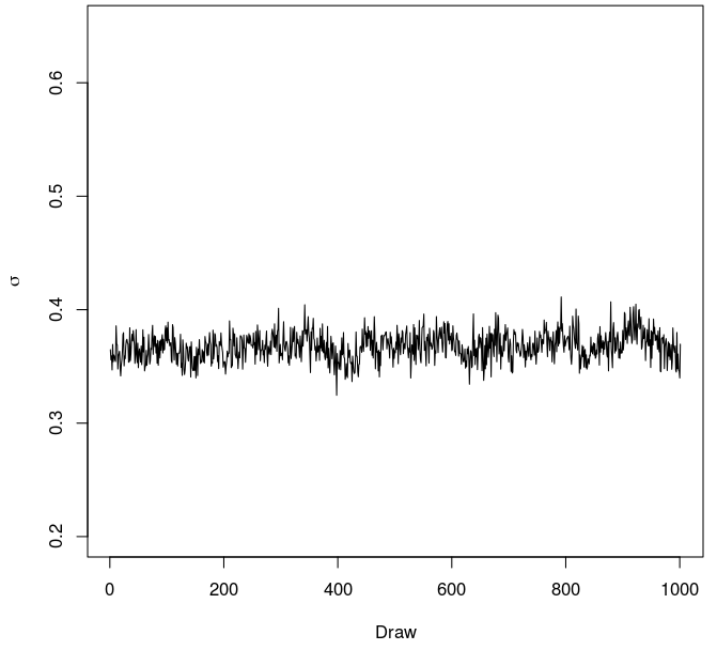

Figure D.15: *PS-DART model (Friedman Function under Sparsity) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

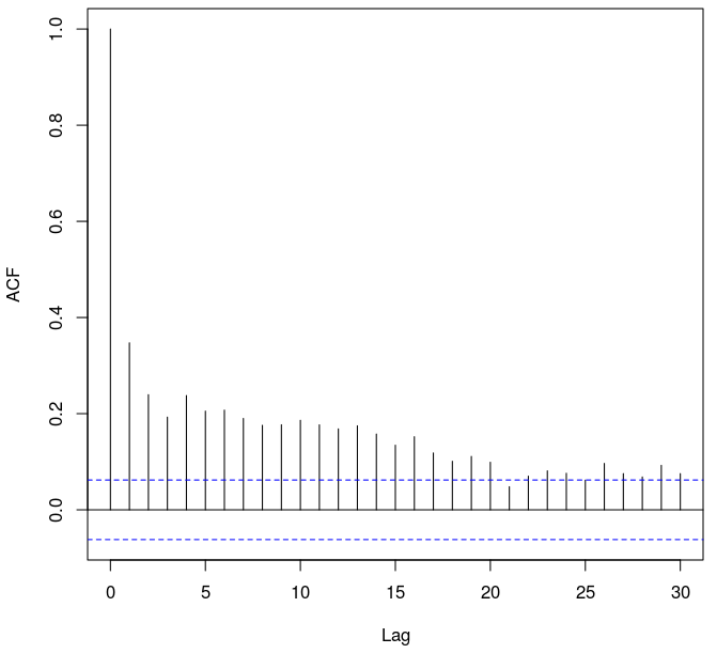

Figure D.16: *PS-DART model (Friedman Function under Sparsity) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

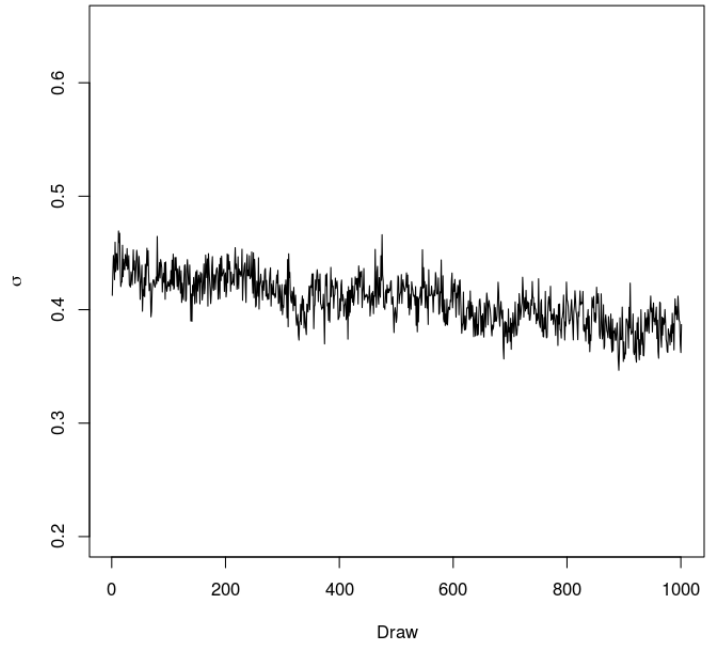

Figure D.17: *GLM-DART model (Friedman Function under Sparsity) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

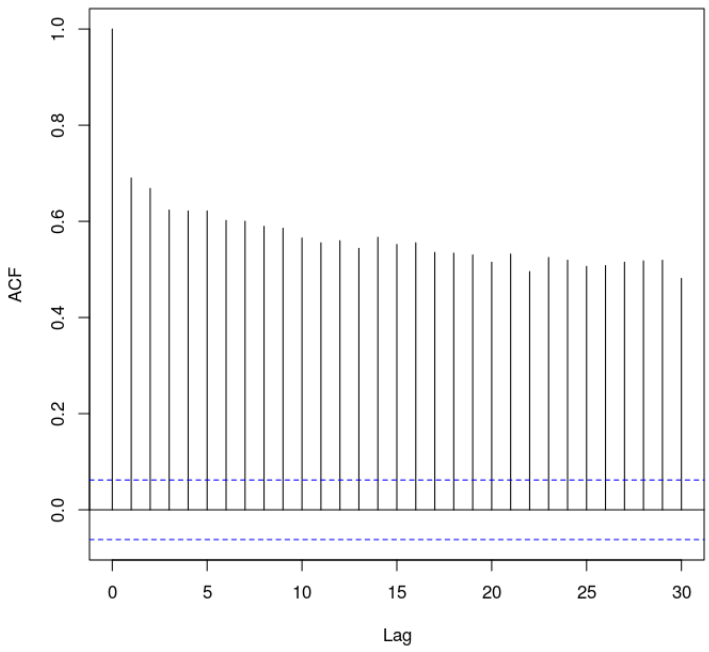

Figure D.18: *GLM-DART model (Friedman Function under Sparsity) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

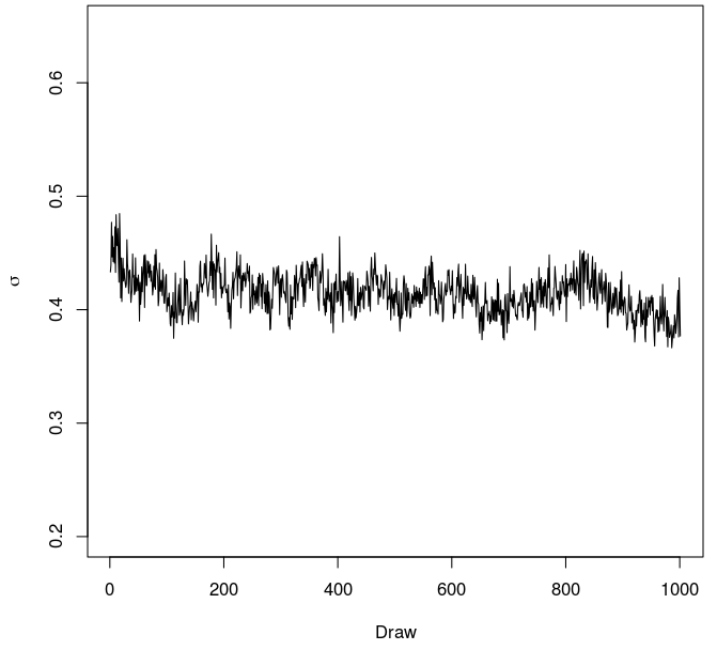

Figure D.19: *Rand-DART model (Friedman Function under Sparsity) - σ posterior draws trace plot. Apparently the draws traverse the sample space with minor oscillations.*

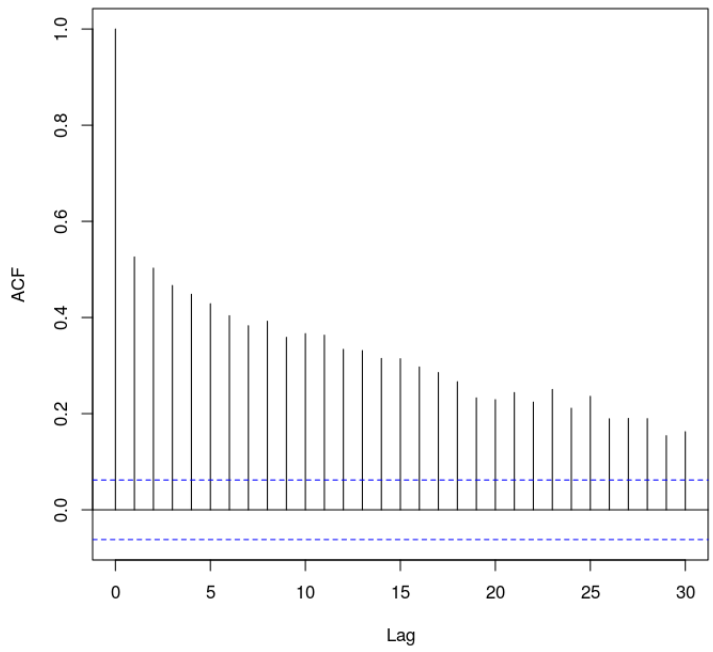

Figure D.20: *Rand-DART model (Friedman Function under Sparsity) - ACF function for the σ draws. Apparently there is some autocorrelation among the draws.*

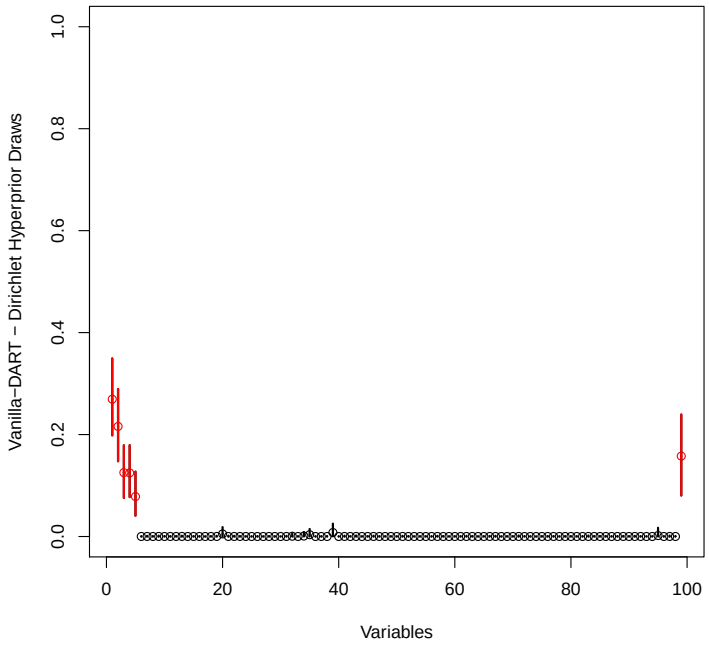

Figure D.21: *Friedman Function under Sparsity - Posterior draws from Vanilla-DART Dirichlet hyperprior with 95% credible intervals. In red:* x_1, x_2, x_3, x_4, x_5, *and* z, *respectively.*

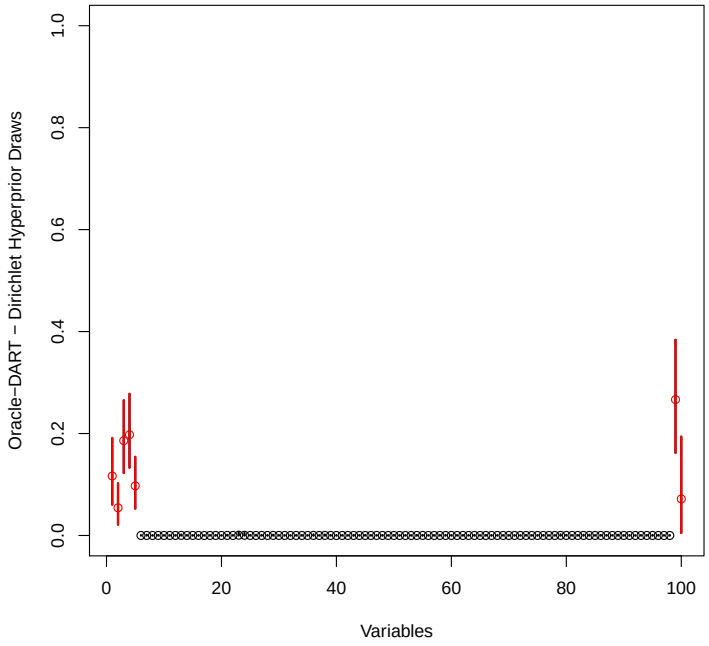

Figure D.22: *Friedman Function under Sparsity - Posterior draws from Oracle-DART Dirichlet hyperprior with 95% credible intervals. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

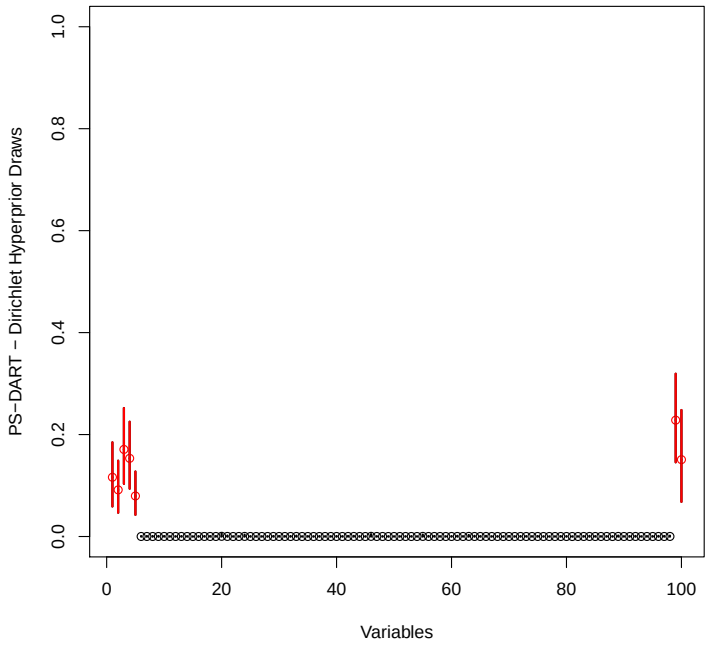

Figure D.23: *Friedman Function under Sparsity - Posterior draws from Oracle-DART Dirichlet hyperprior with 95% credible intervals. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

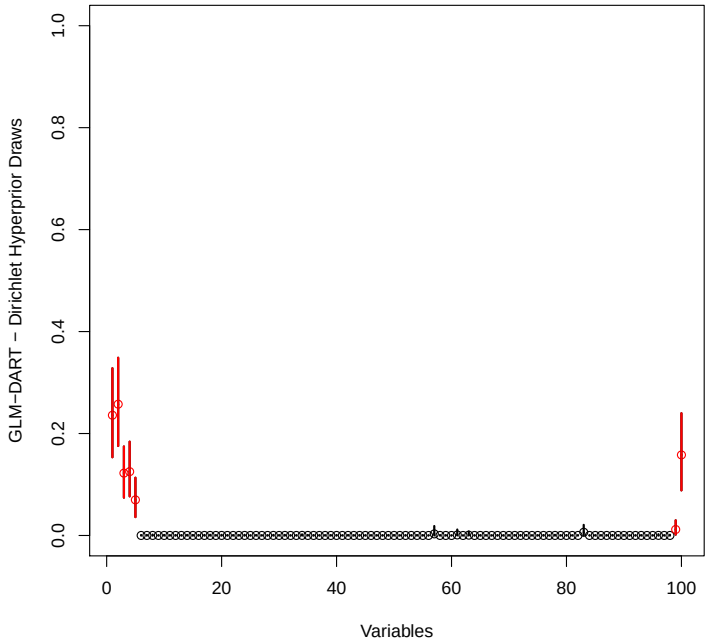

Figure D.24: *Friedman Function under Sparsity - Posterior draws from GLM-DART Dirichlet hyperprior with 95% credible intervals. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

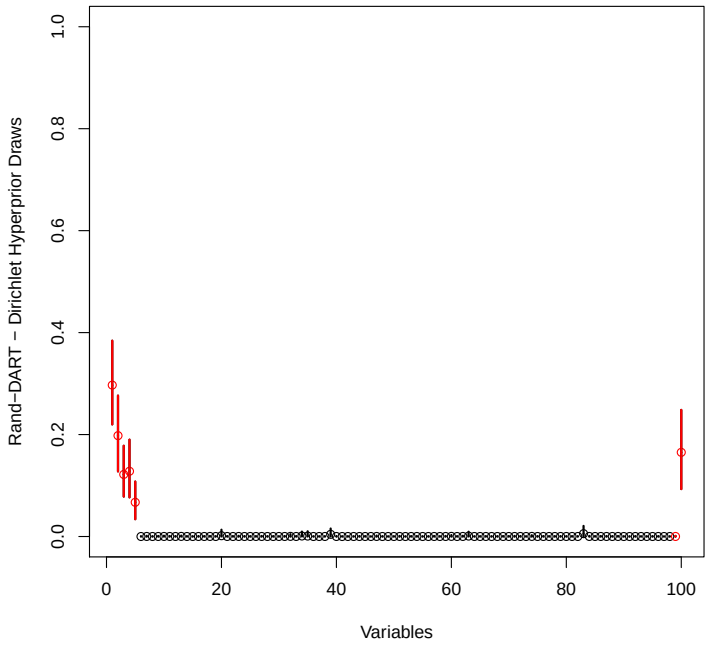

Figure D.25: *Friedman Function under Sparsity - Posterior draws from Rand-DART Dirichlet hyperprior with 95% credible intervals. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

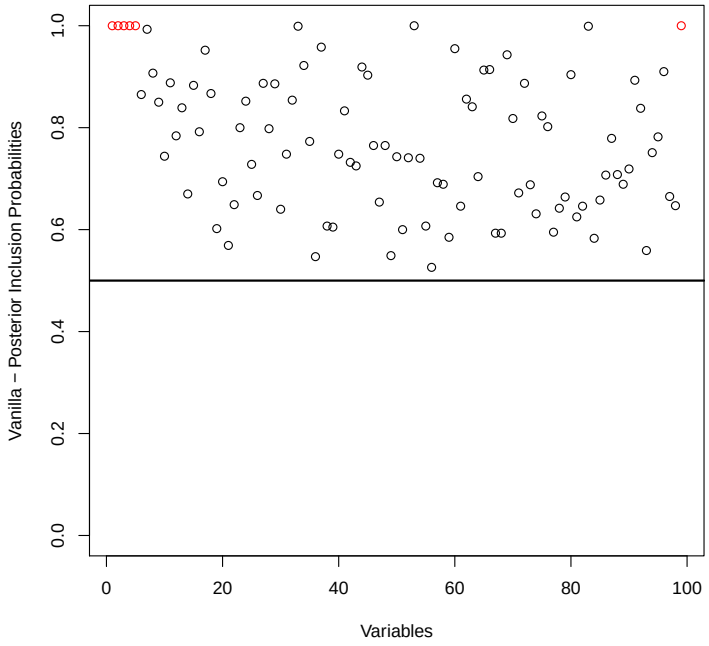

Figure D.26: *Friedman Function under Sparsity - Posterior Inclusion Probability of Vanilla model. In red:* x_1, x_2, x_3, x_4, x_5, *and* z, *respectively.*

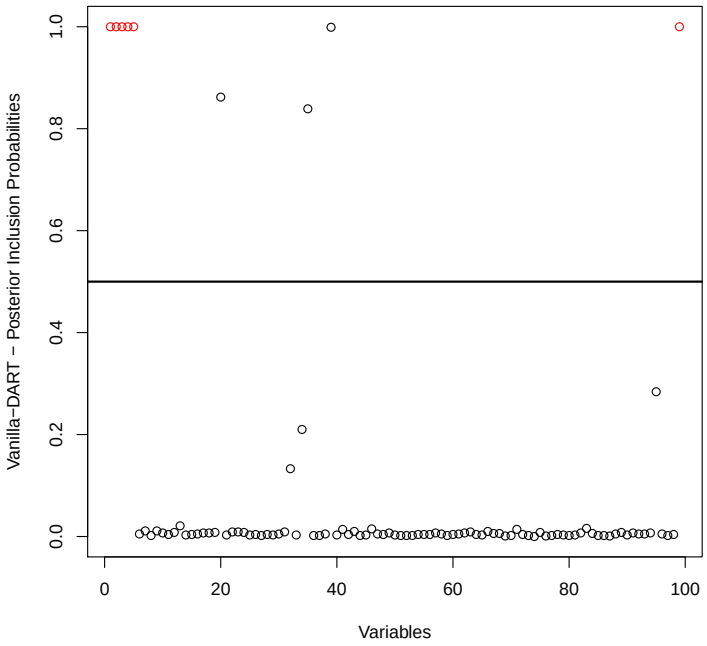

Figure D.27: *Friedman Function under Sparsity - Posterior Inclusion Probability of Vanilla-DART model. In red:* x_1, x_2, x_3, x_4, x_5, *and* z, *respectively.*

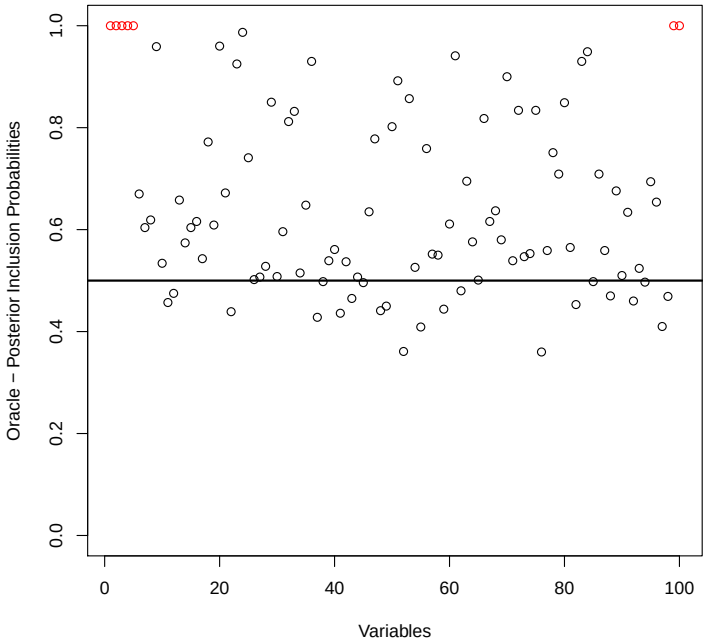

Figure D.28: *Friedman Function under Sparsity - Posterior Inclusion Probability of Oracle model. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

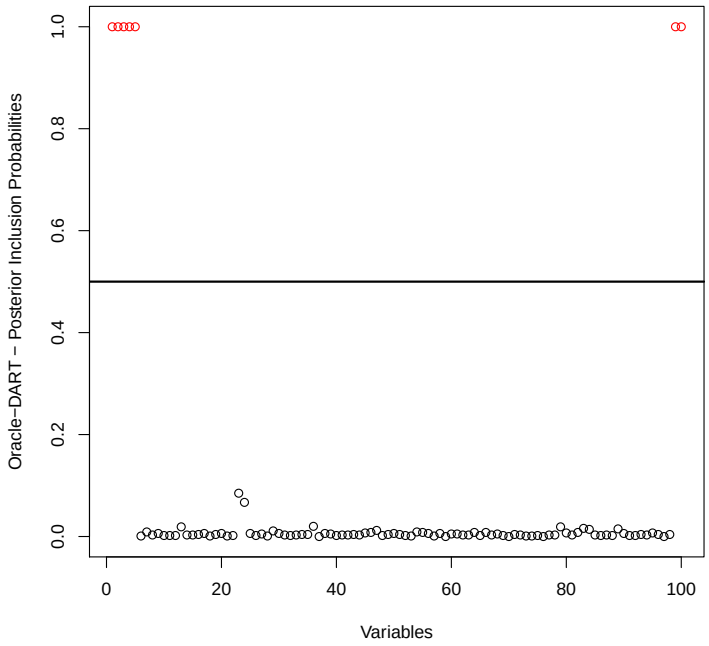

Figure D.29: *Friedman Function under Sparsity - Posterior Inclusion Probability of Oracle-DART model. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

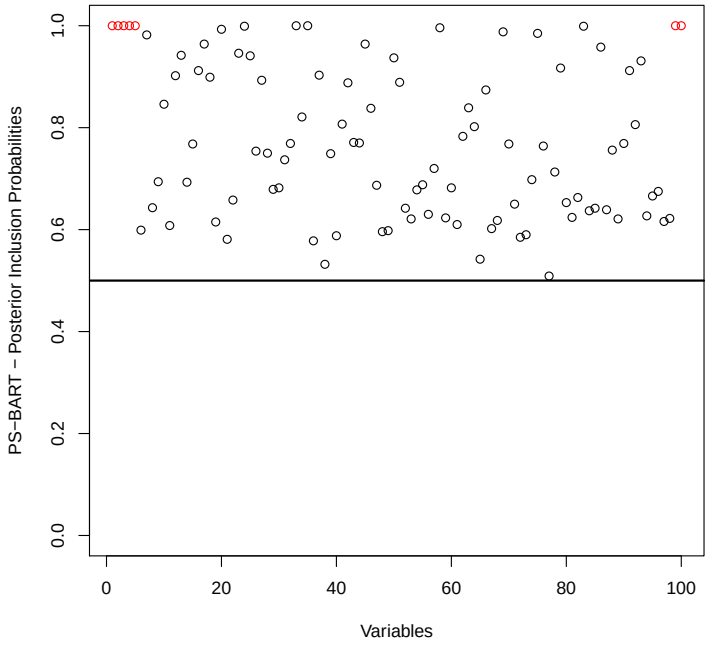

Figure D.30: *Friedman Function under Sparsity - Posterior Inclusion Probability of PS-BART model. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

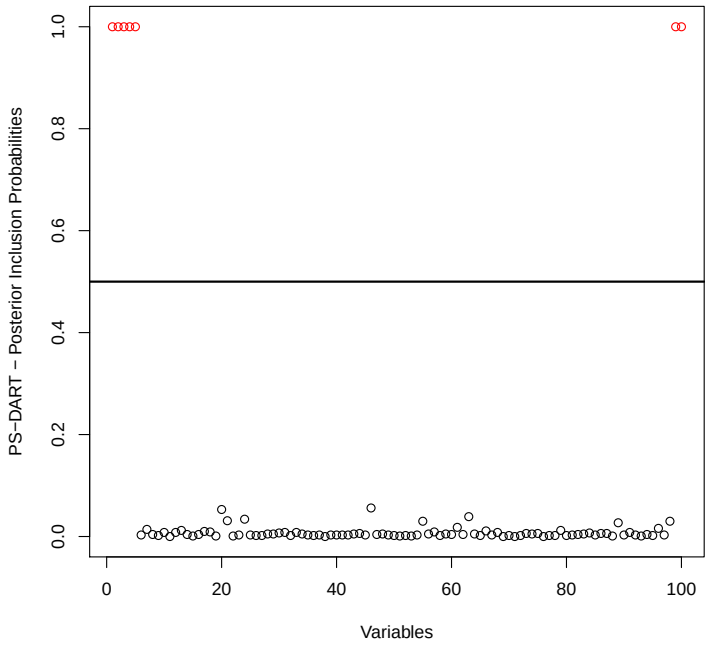

Figure D.31: *Friedman Function under Sparsity - Posterior Inclusion Probability of PS-DART model. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

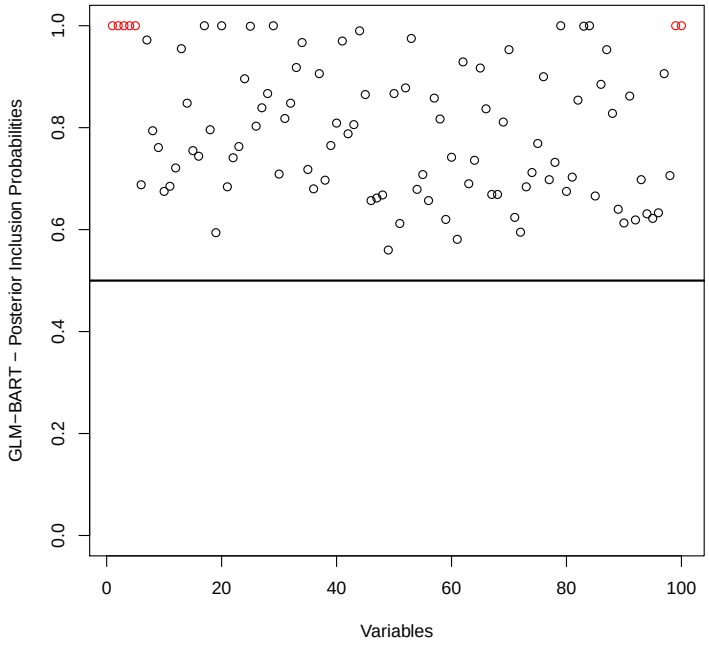

Figure D.32: *Friedman Function under Sparsity - Posterior Inclusion Probability of GLM-BART model. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

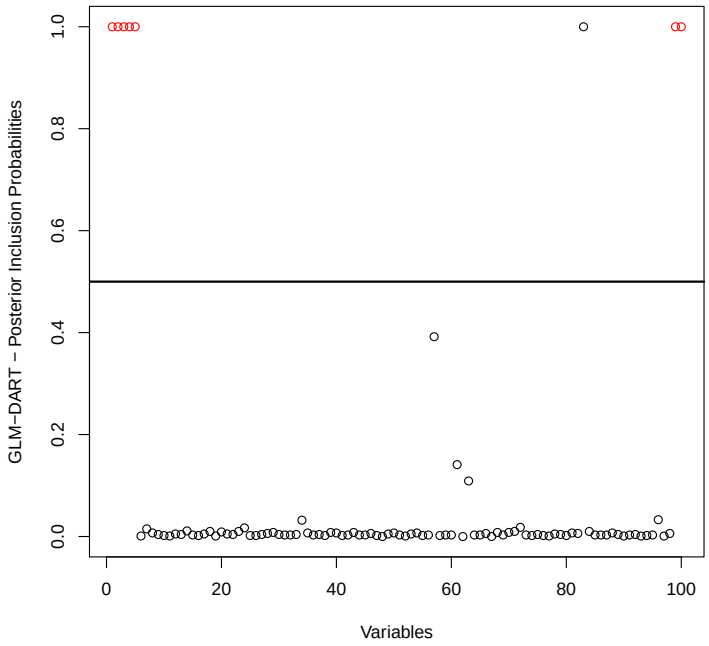

Figure D.33: *Friedman Function under Sparsity - Posterior Inclusion Probability of GLM-DART model. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

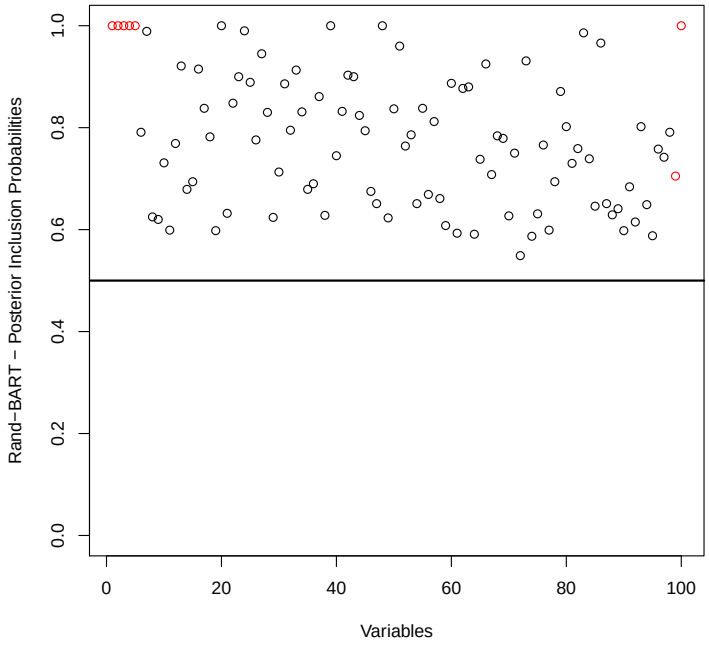

Figure D.34: *Friedman Function under Sparsity - Posterior Inclusion Probability of Rand-BART model. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

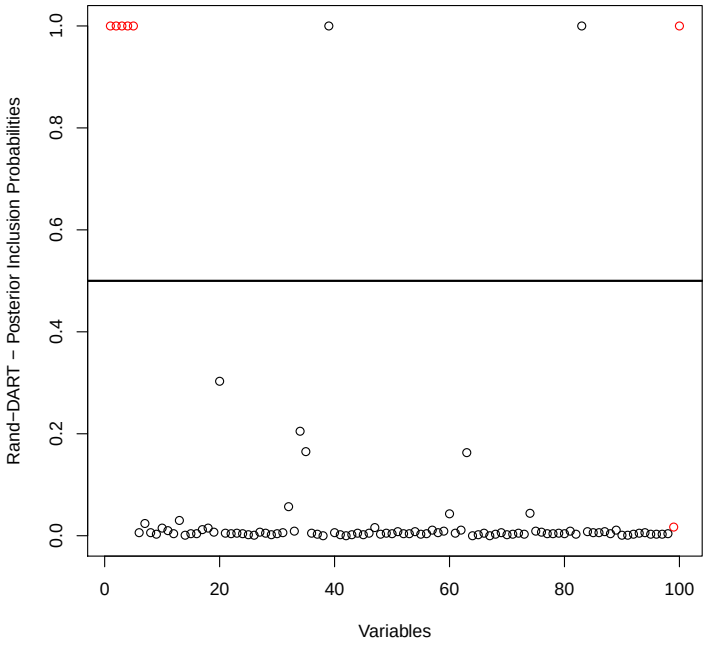

Figure D.35: *Friedman Function under Sparsity - Posterior Inclusion Probability of Rand-DART model. In red:* x_1, x_2, x_3, x_4, x_5, $\pi(\tilde{x})$ *and* z, *respectively.*

D.2 ICE Plots

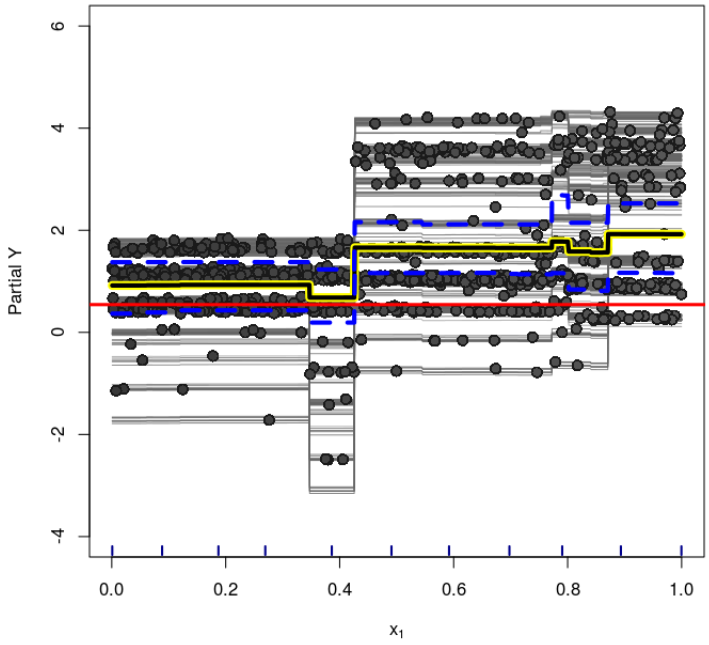

Figure D.36: *Vanilla model (Friedman Function under Sparsity) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

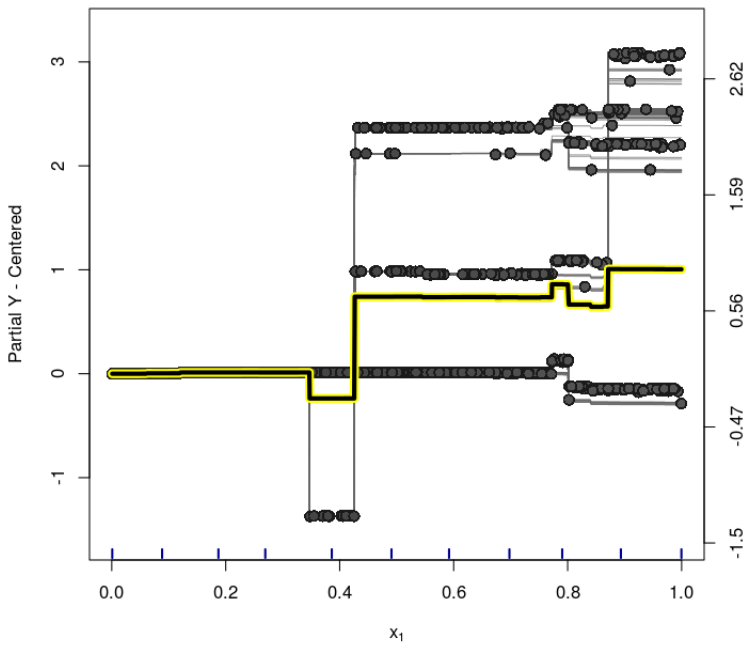

Figure D.37: *Vanilla model (Friedman Function under Sparsity) - Variable x_1 - centered-ICE Plot for the treatment effect.*

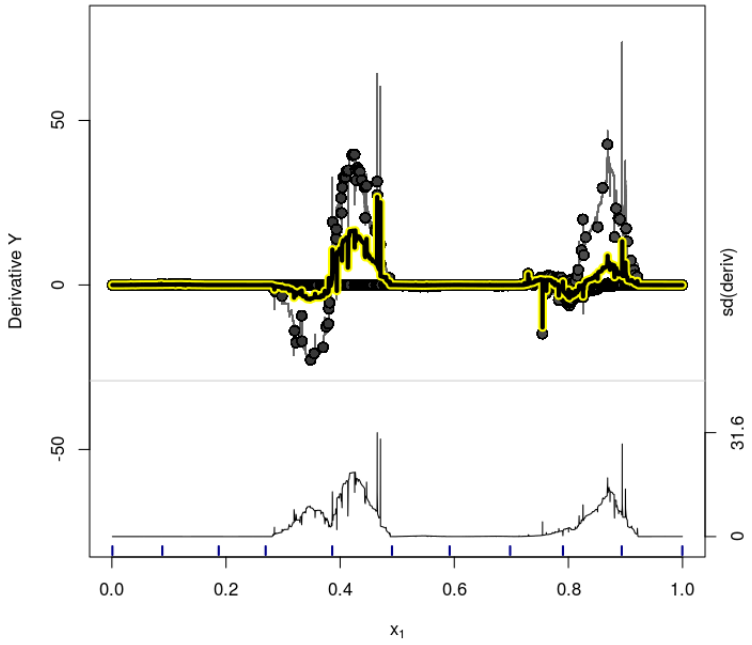

Figure D.38: *Vanilla model (Friedman Function under Sparsity) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

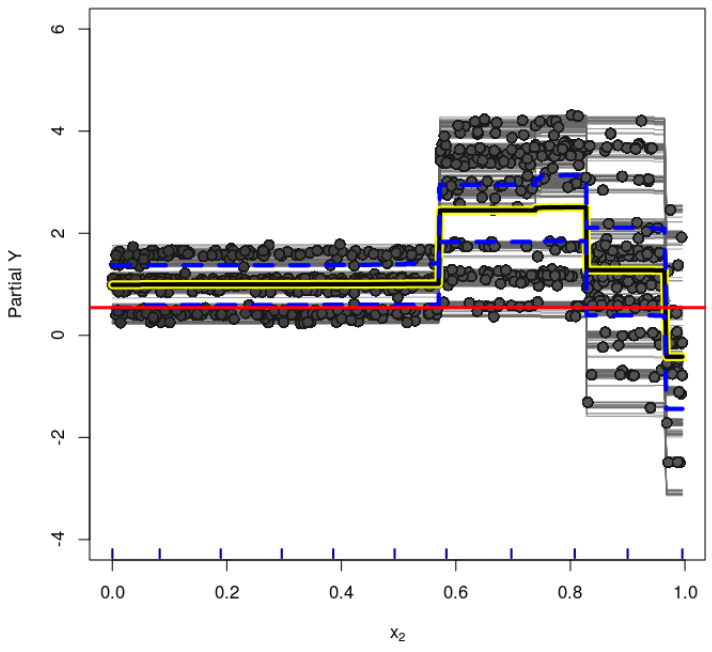

Figure D.39: *Vanilla model (Friedman Function under Sparsity) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

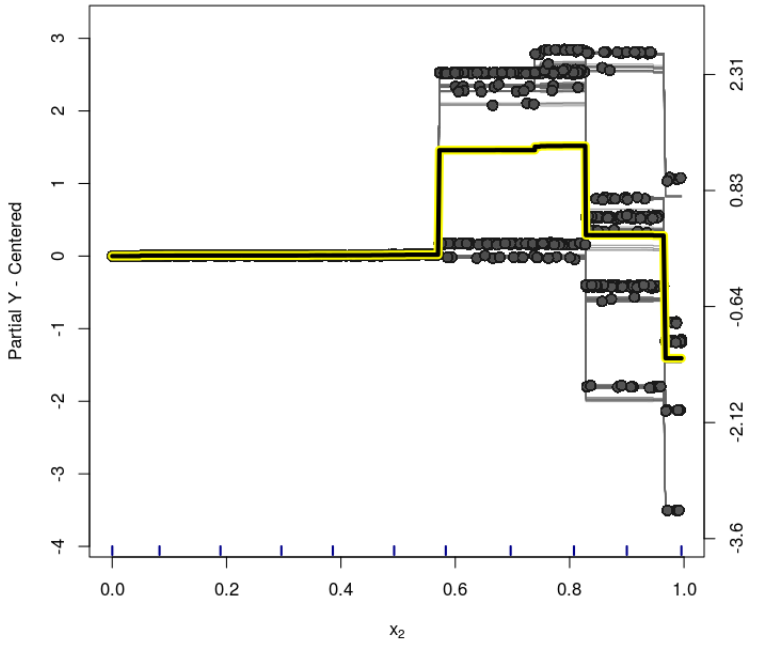

Figure D.40: *Vanilla model (Friedman Function under Sparsity) - Variable x_2 - centered-ICE Plot for the treatment effect.*

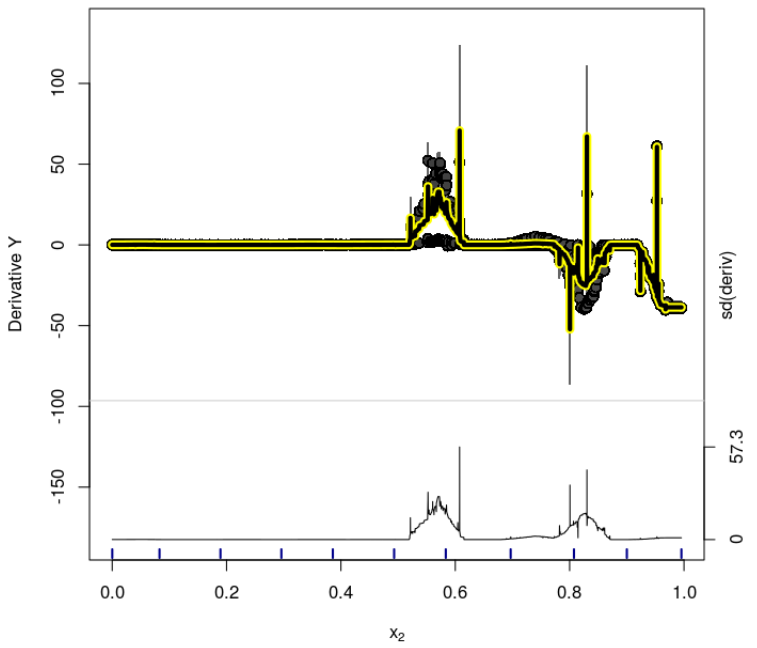

Figure D.41: *Vanilla model (Friedman Function under Sparsity) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

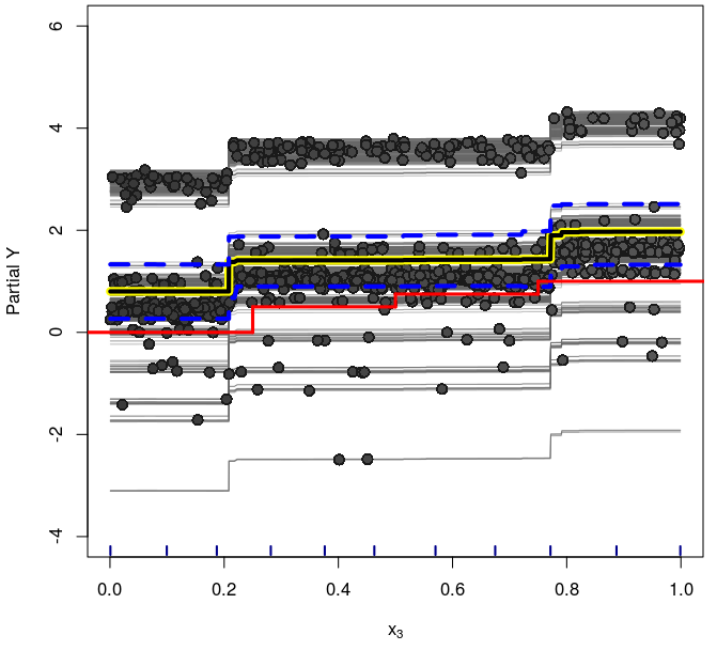

Figure D.42: *Vanilla model (Friedman Function under Sparsity) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

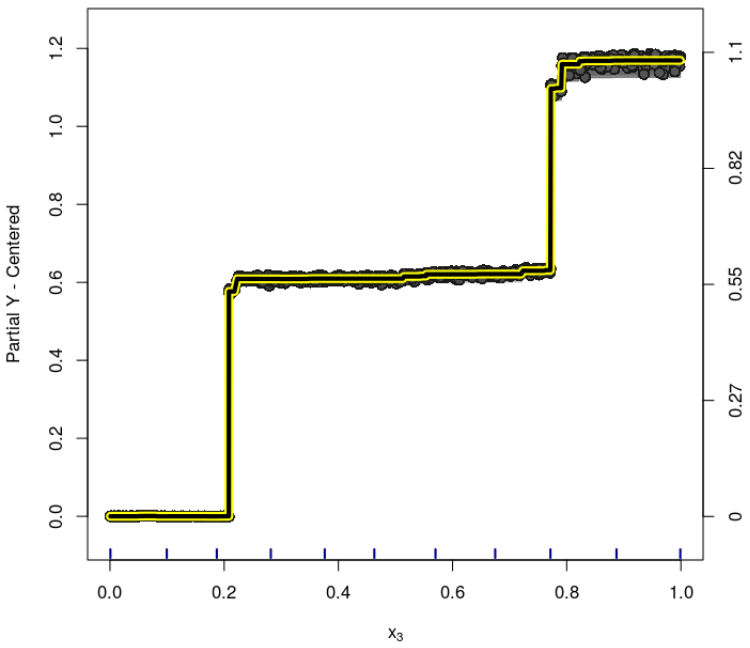

Figure D.43: *Vanilla model (Friedman Function under Sparsity) - Variable x_3 - centered-ICE Plot for the treatment effect.*

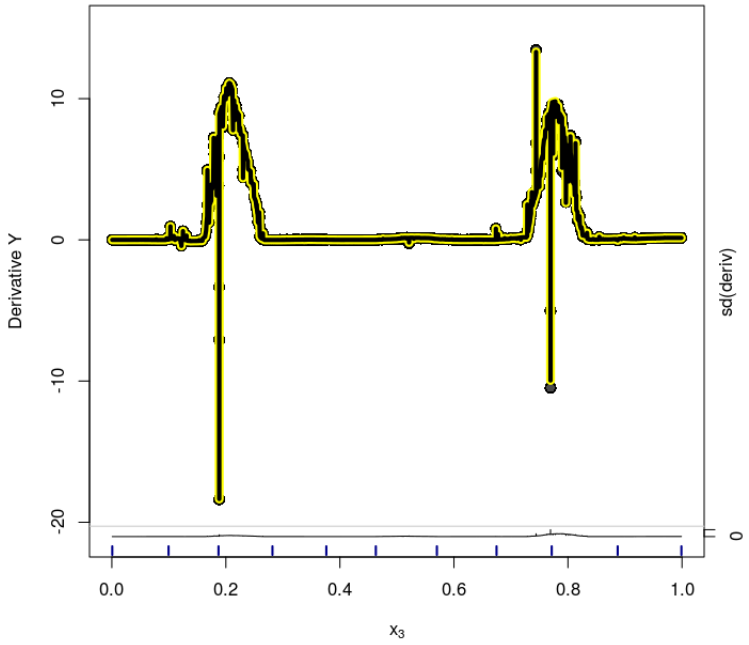

Figure D.44: *Vanilla model (Friedman Function under Sparsity) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

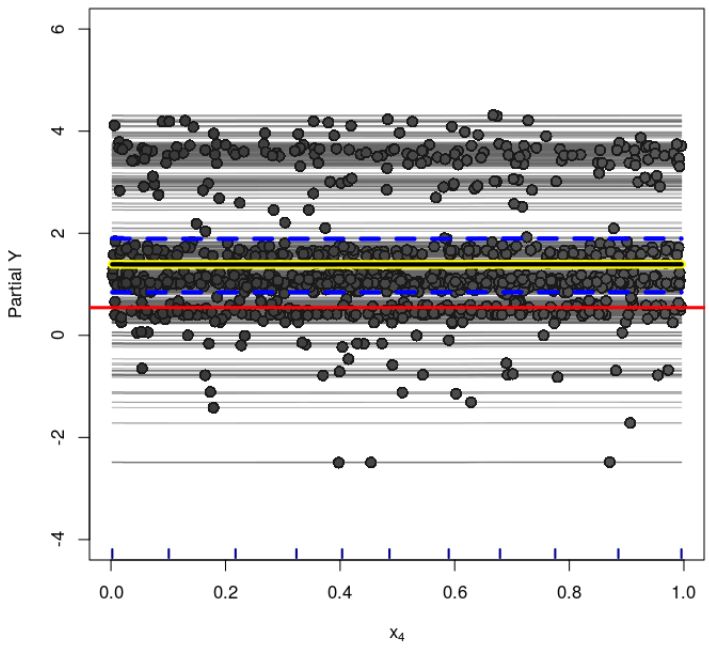

Figure D.45: *Vanilla model (Friedman Function under Sparsity) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

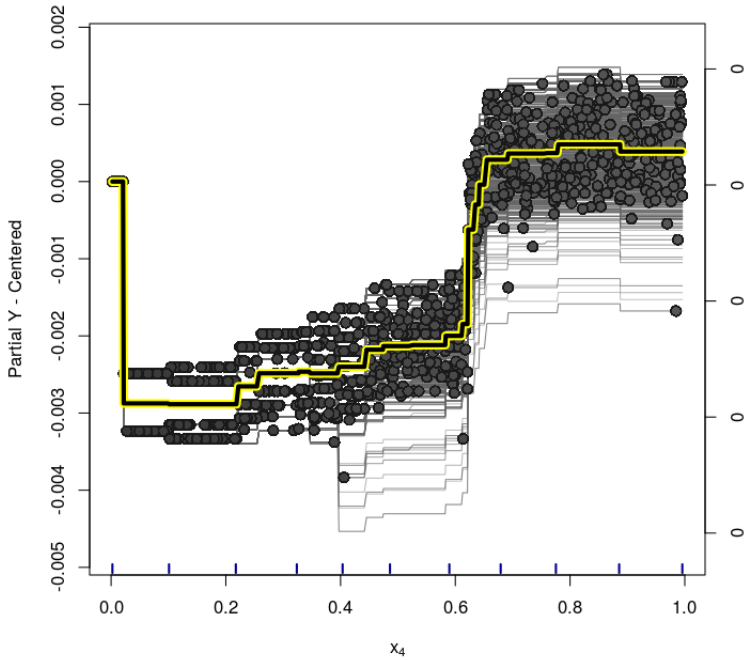

Figure D.46: *Vanilla model (Friedman Function under Sparsity) - Variable x_4 - centered-ICE Plot for the treatment effect.*

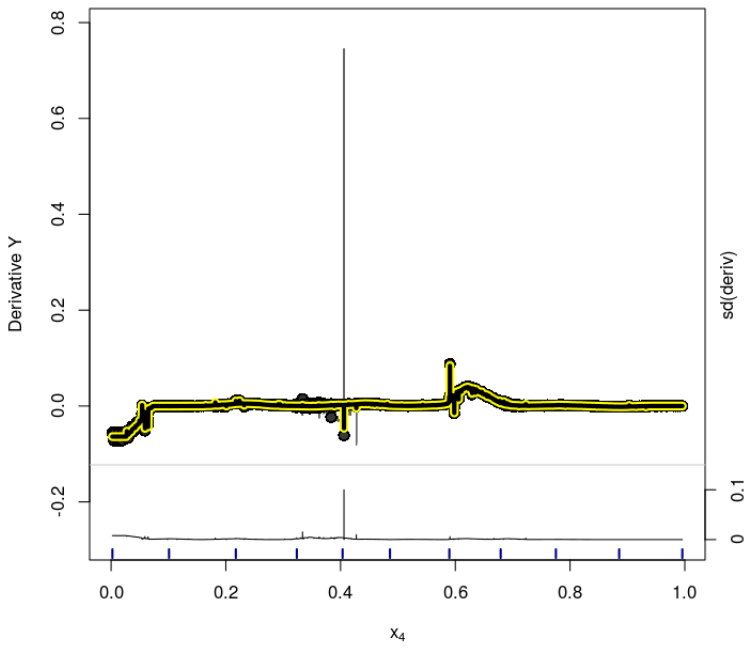

Figure D.47: *Vanilla model (Friedman Function under Sparsity) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

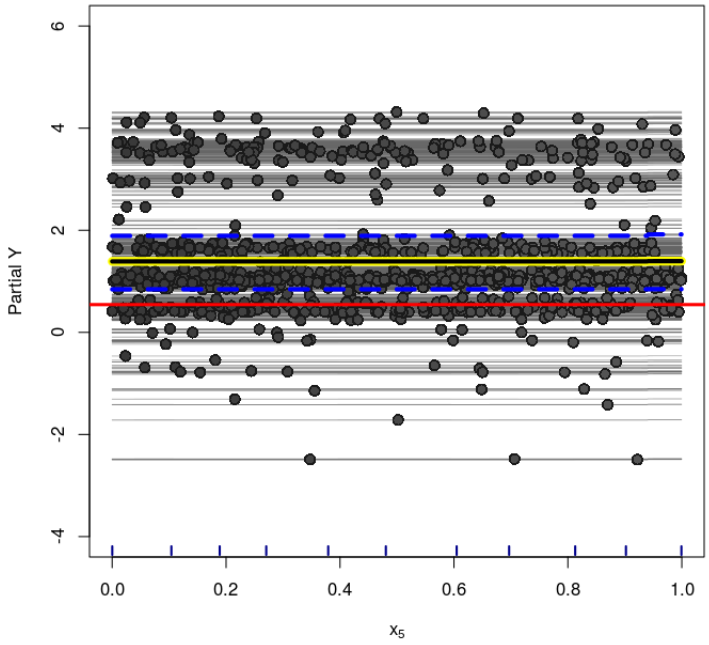

Figure D.48: *Vanilla model (Friedman Function under Sparsity) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

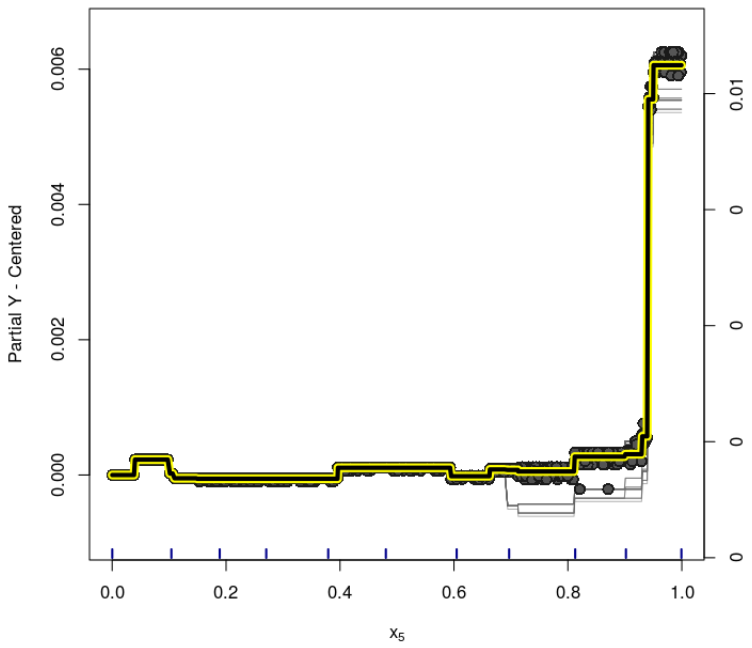

Figure D.49: *Vanilla model (Friedman Function under Sparsity) - Variable x_5 - centered-ICE Plot for the treatment effect.*

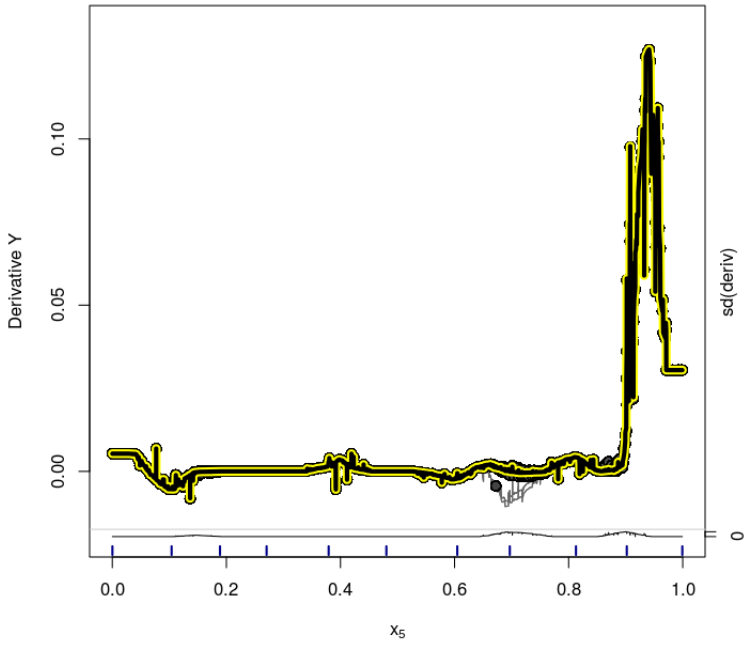

Figure D.50: *Vanilla model (Friedman Function under Sparsity) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

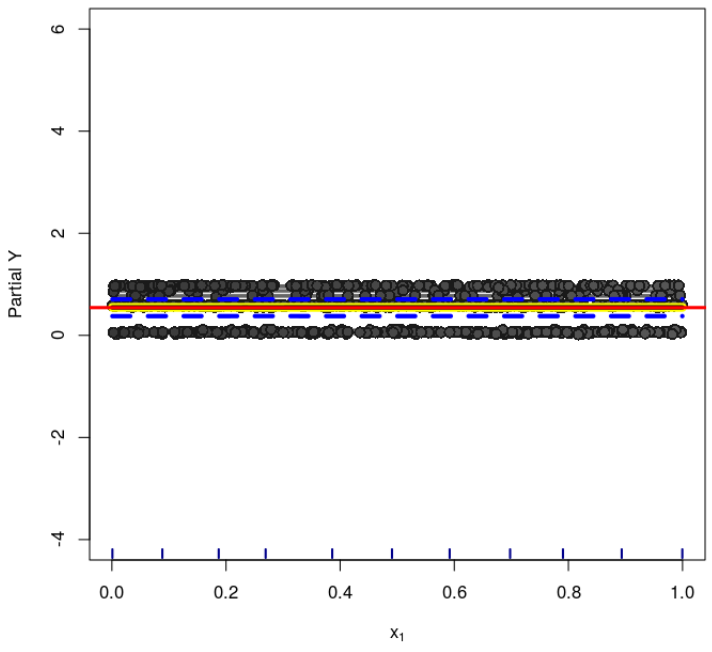

Figure D.51: *Oracle model (Friedman Function under Sparsity) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

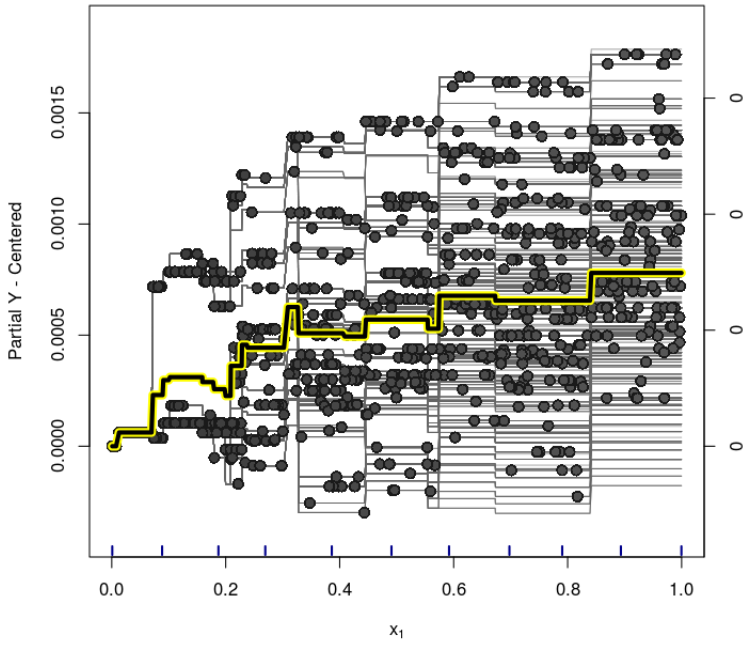

Figure D.52: *Oracle model (Friedman Function under Sparsity) - Variable x_1 - centered-ICE Plot for the treatment effect.*

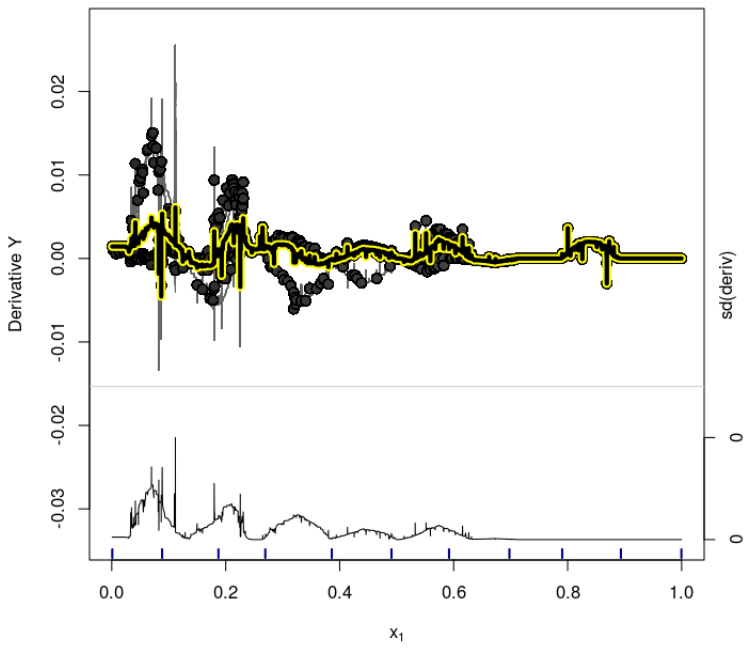

Figure D.53: *Oracle model (Friedman Function under Sparsity) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

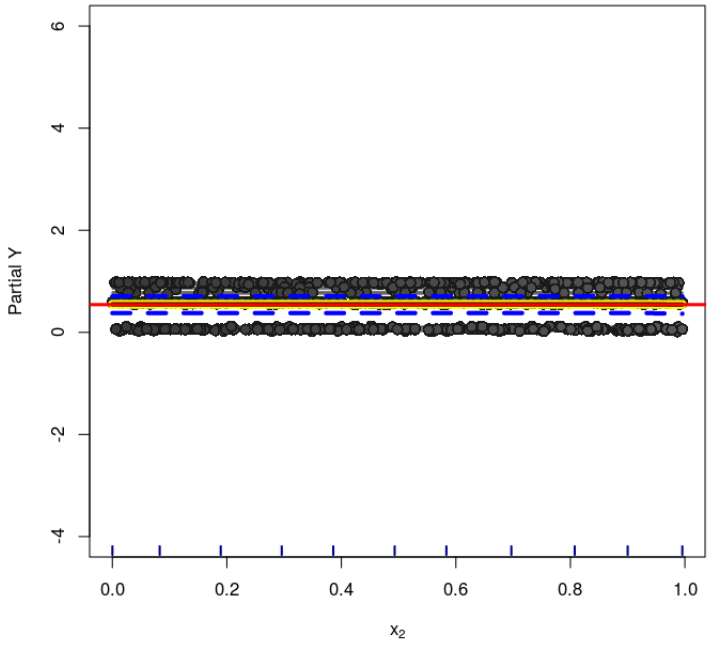

Figure D.54: *Oracle model (Friedman Function under Sparsity) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

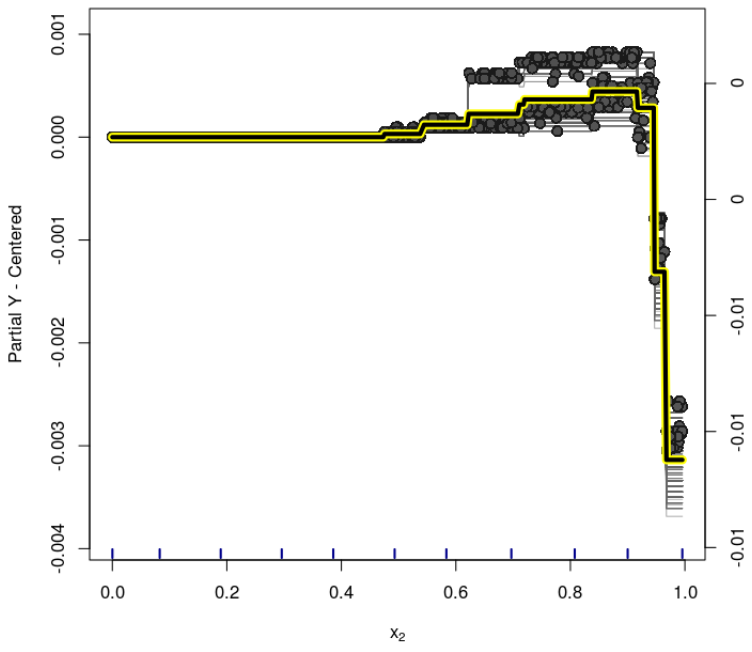

Figure D.55: *Oracle model (Friedman Function under Sparsity) - Variable x_2 - centered-ICE Plot for the treatment effect.*

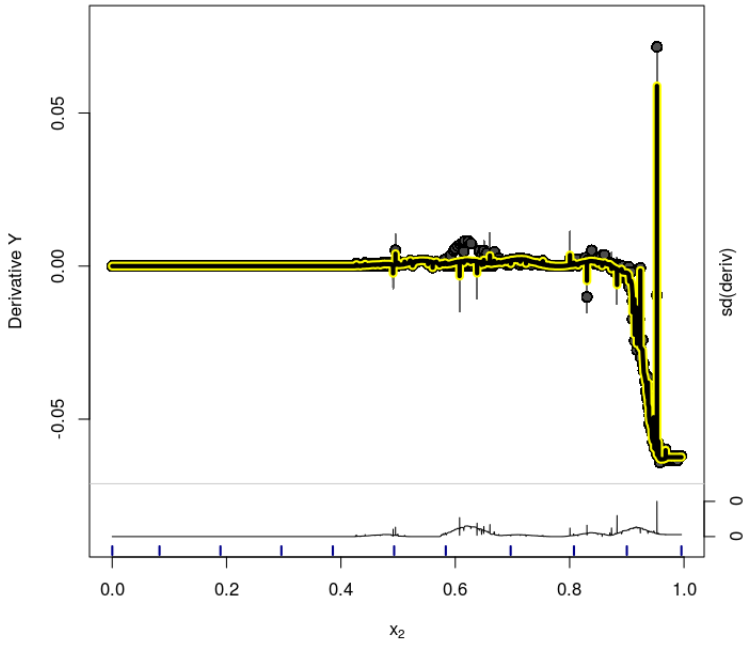

Figure D.56: *Oracle model (Friedman Function under Sparsity) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

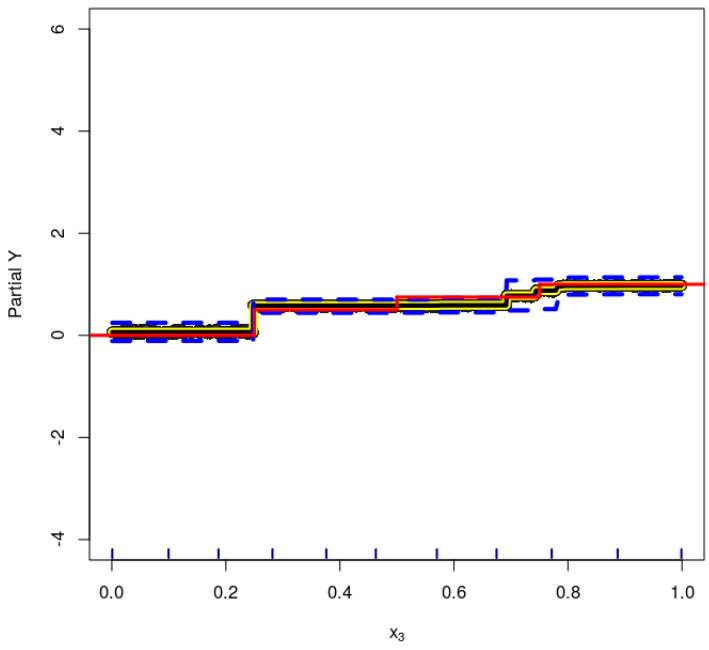

Figure D.57: *Oracle model (Friedman Function under Sparsity) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

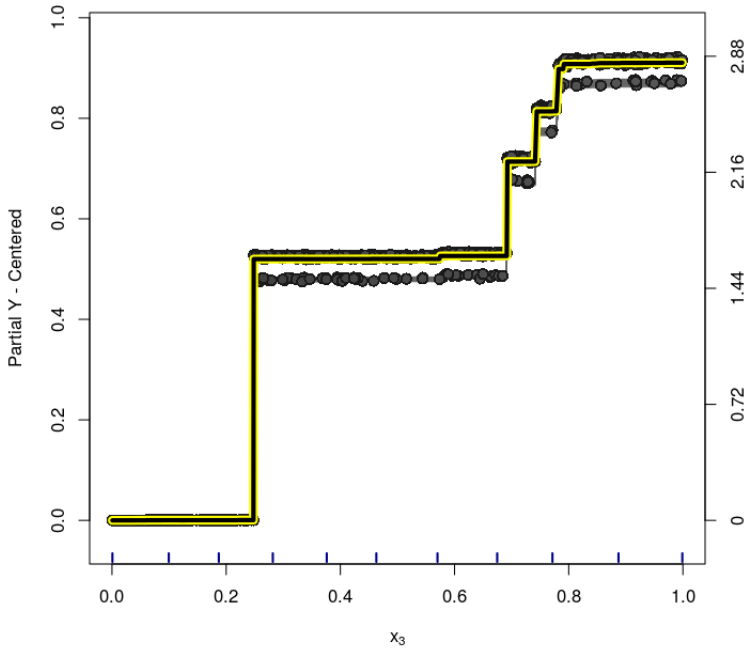

Figure D.58: *Oracle model (Friedman Function under Sparsity) - Variable x_3 - centered-ICE Plot for the treatment effect.*

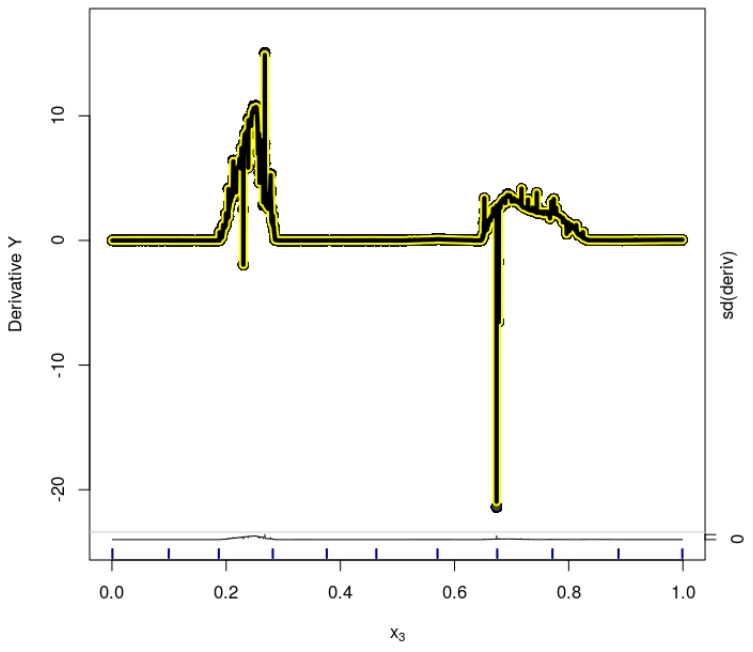

Figure D.59: *Oracle model (Friedman Function under Sparsity) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

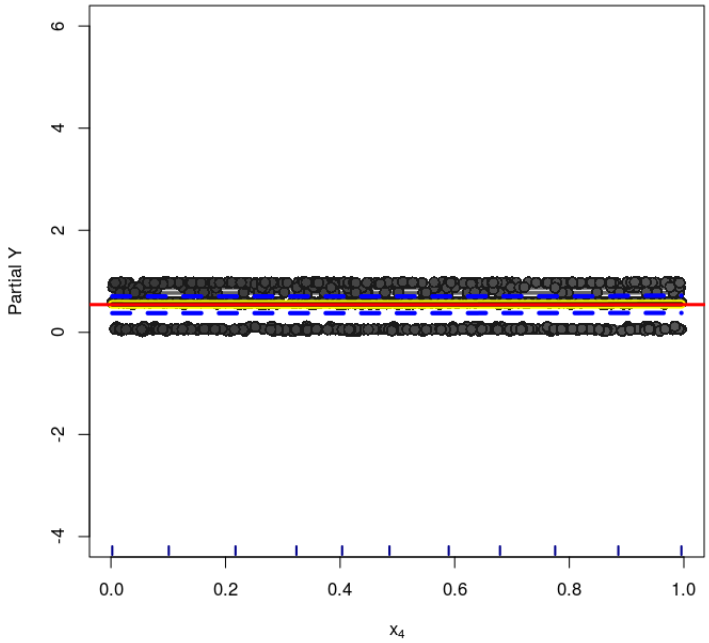

Figure D.60: *Oracle model (Friedman Function under Sparsity) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

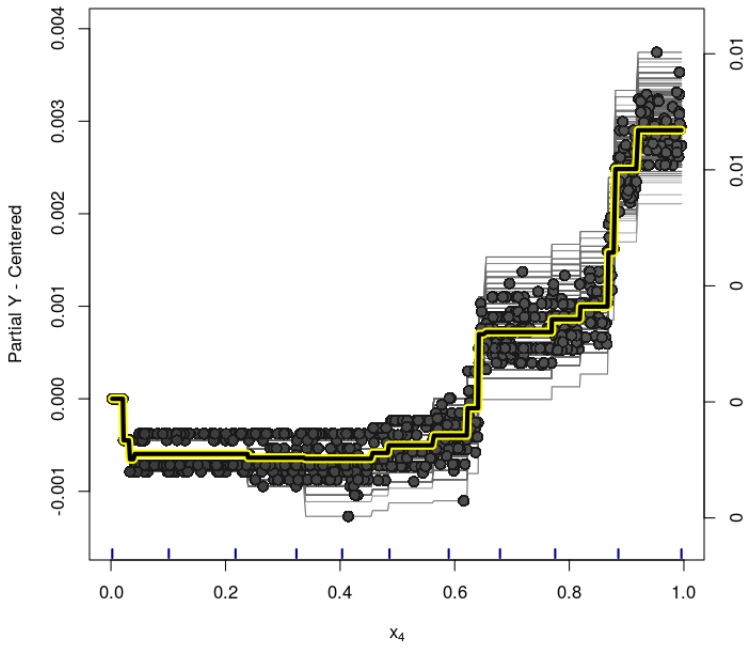

Figure D.61: *Oracle model (Friedman Function under Sparsity) - Variable x_4 - centered-ICE Plot for the treatment effect.*

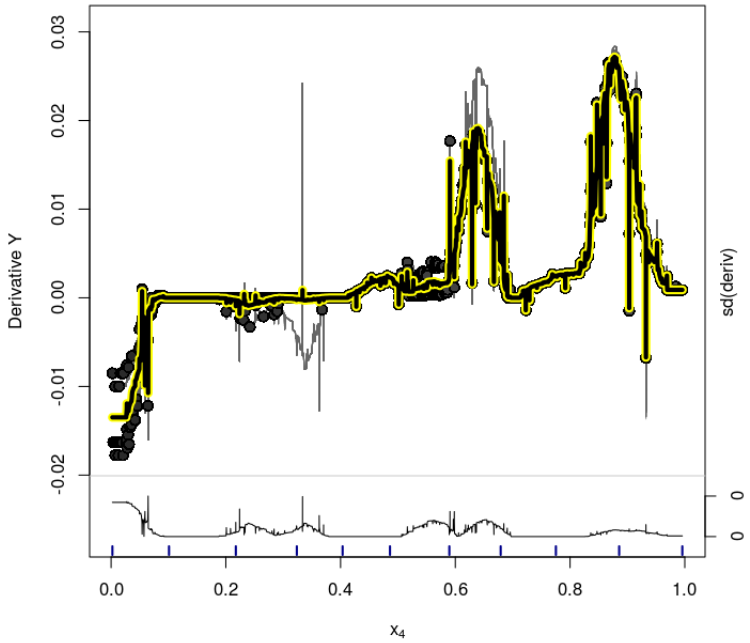

Figure D.62: *Oracle model (Friedman Function under Sparsity) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

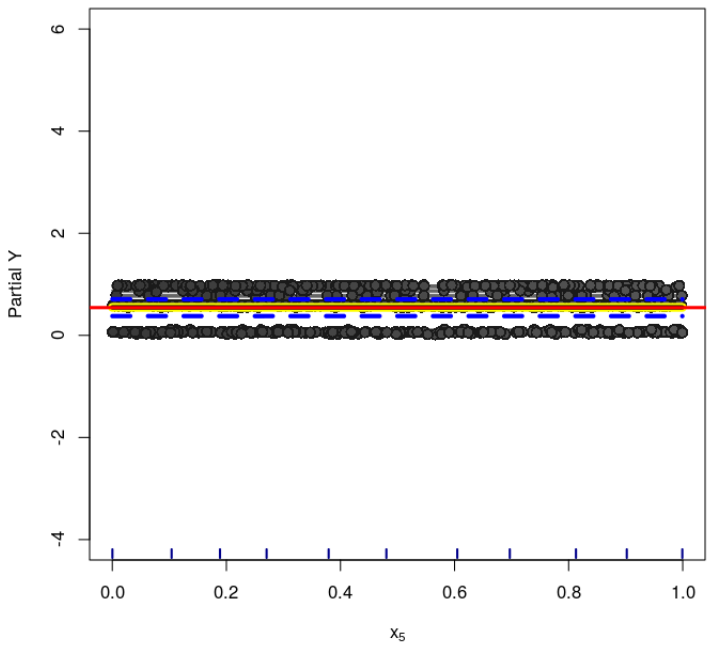

Figure D.63: *Oracle model (Friedman Function under Sparsity) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

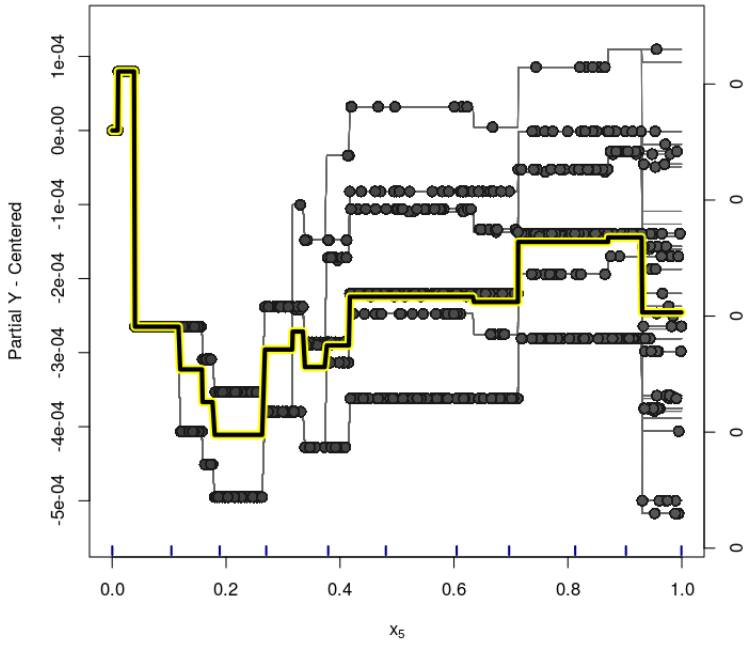

Figure D.64: *Oracle model (Friedman Function under Sparsity) - Variable x_5 - centered-ICE Plot for the treatment effect.*

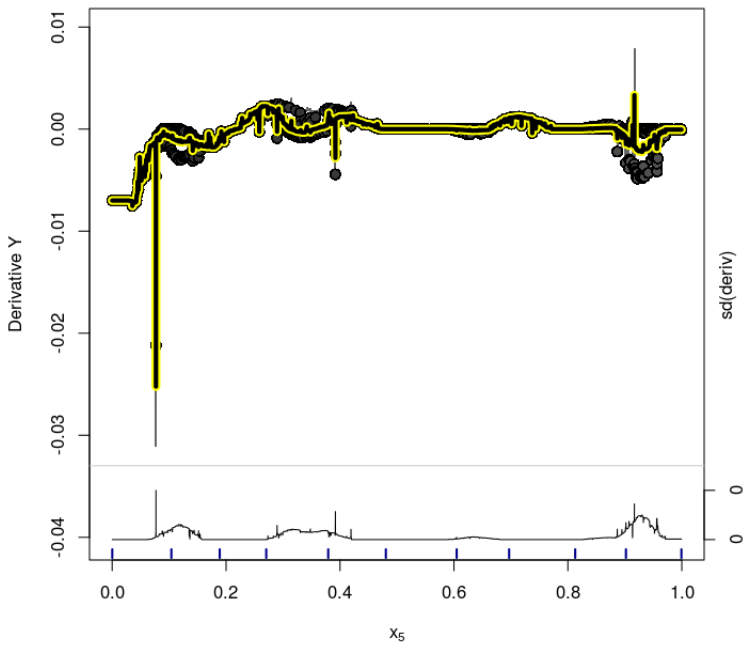

Figure D.65: *Oracle model (Friedman Function under Sparsity) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

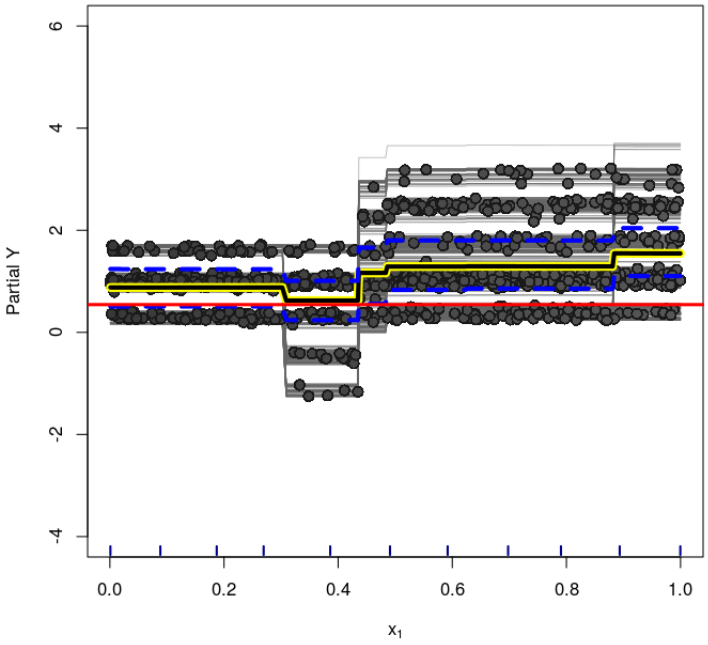

Figure D.66: *PS-BART model (Friedman Function under Sparsity) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

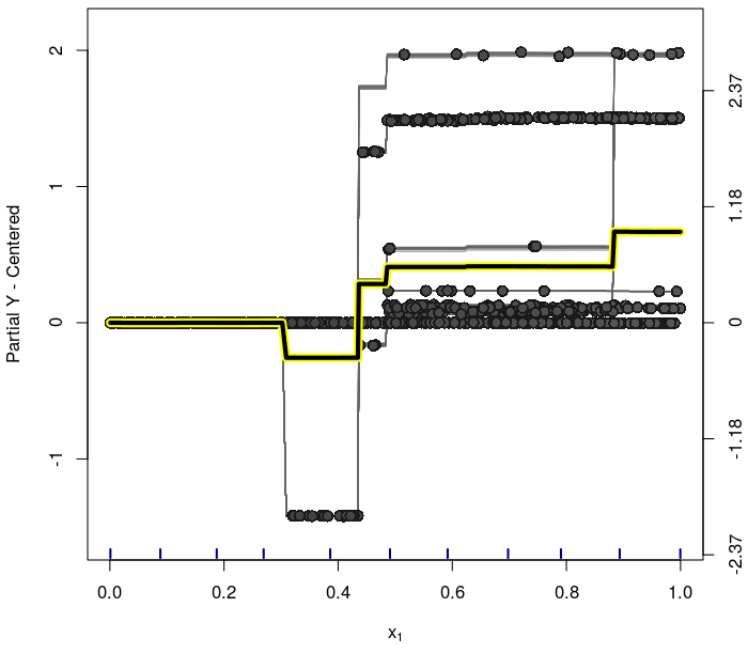

Figure D.67: *PS-BART model (Friedman Function under Sparsity) - Variable x_1 - centered-ICE Plot for the treatment effect.*

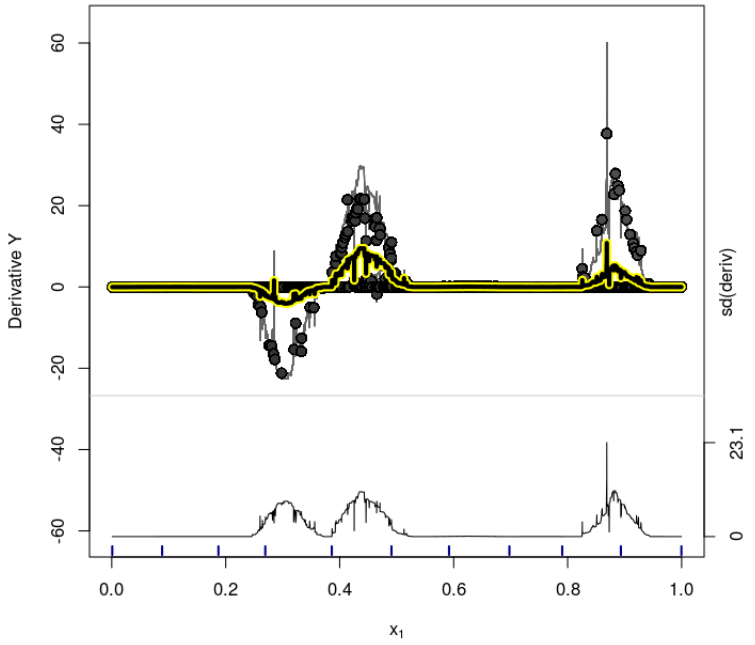

Figure D.68: *PS-BART model (Friedman Function under Sparsity) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

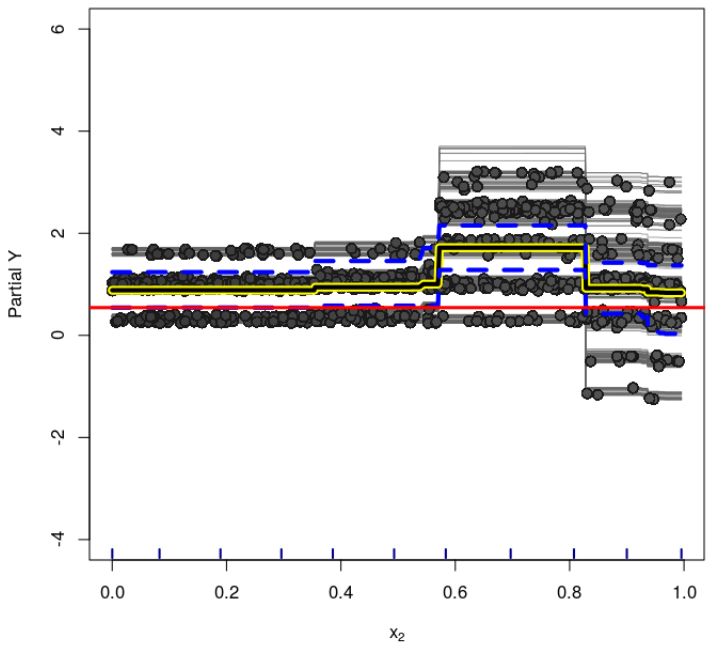

Figure D.69: *PS-BART model (Friedman Function under Sparsity) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

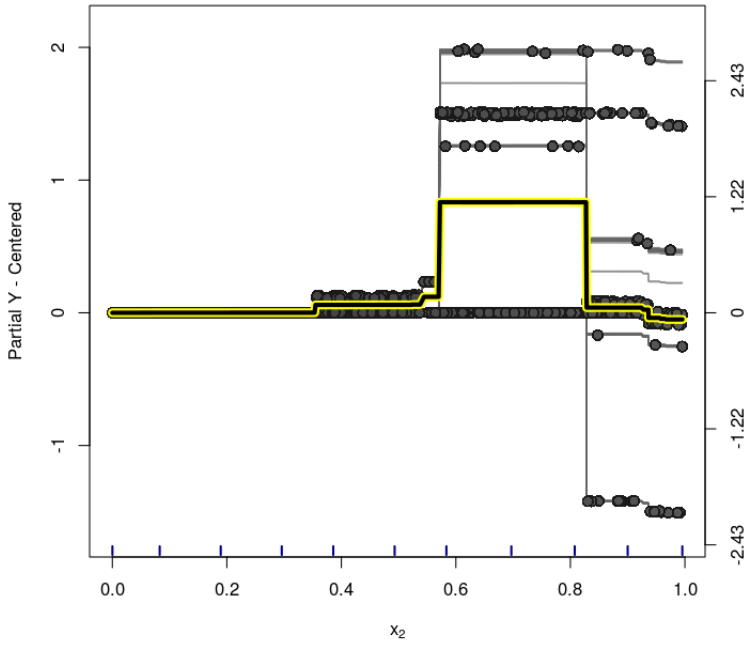

Figure D.70: *PS-BART model (Friedman Function under Sparsity) - Variable x_2 - centered-ICE Plot for the treatment effect.*

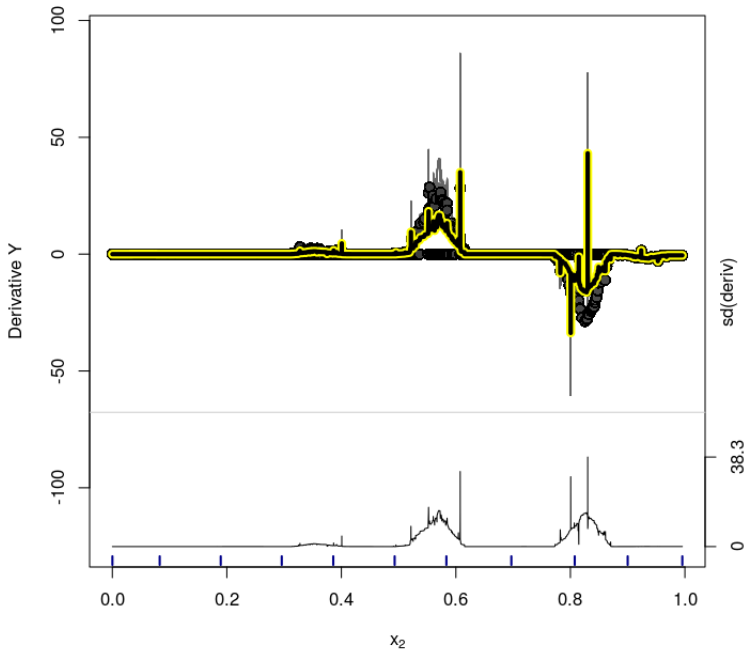

Figure D.71: *PS-BART model (Friedman Function under Sparsity) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

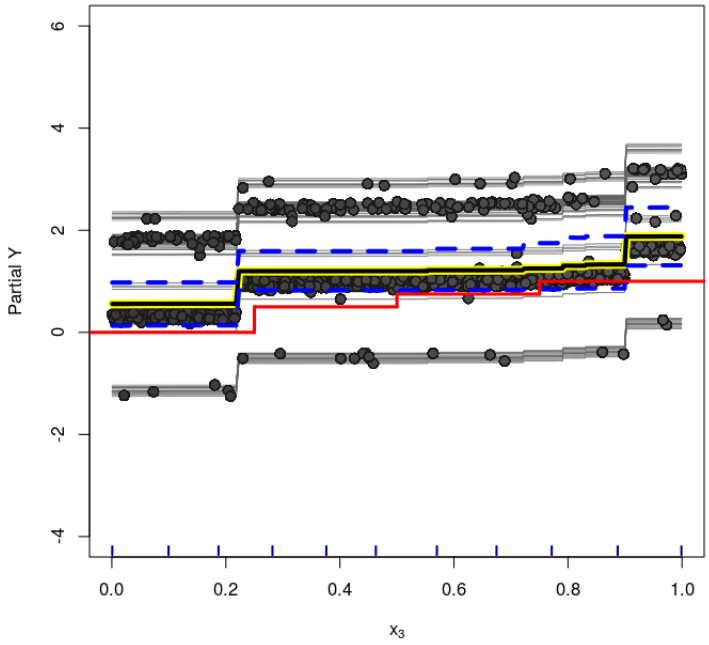

Figure D.72: *PS-BART model (Friedman Function under Sparsity) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

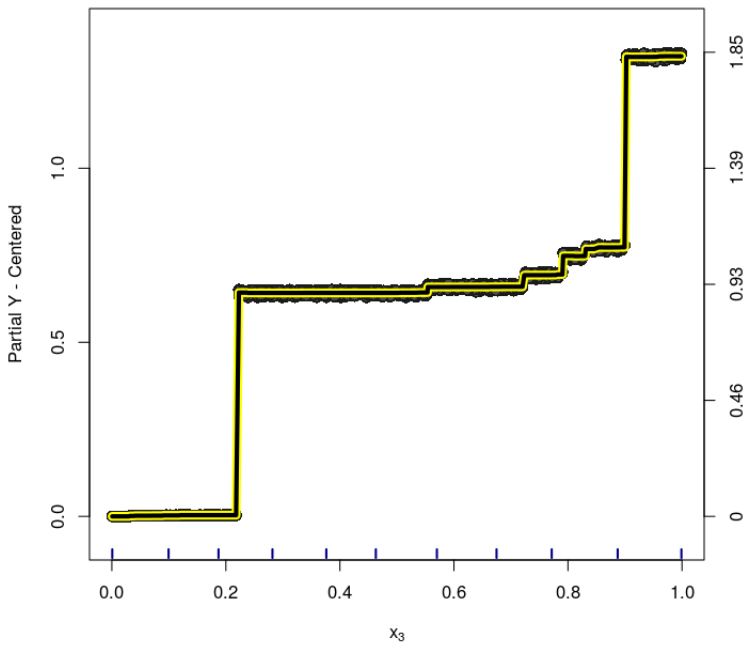

Figure D.73: *PS-BART model (Friedman Function under Sparsity) - Variable x_3 - centered-ICE Plot for the treatment effect.*

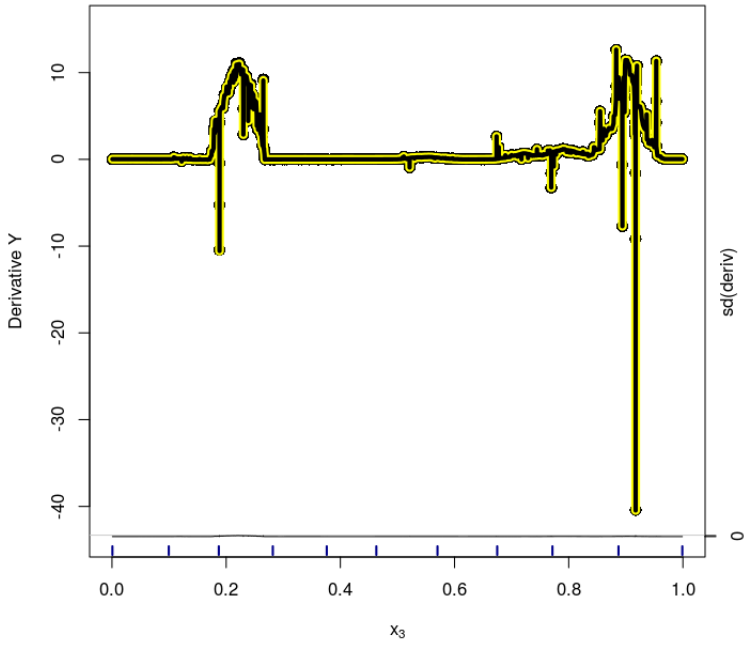

Figure D.74: *PS-BART model (Friedman Function under Sparsity) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

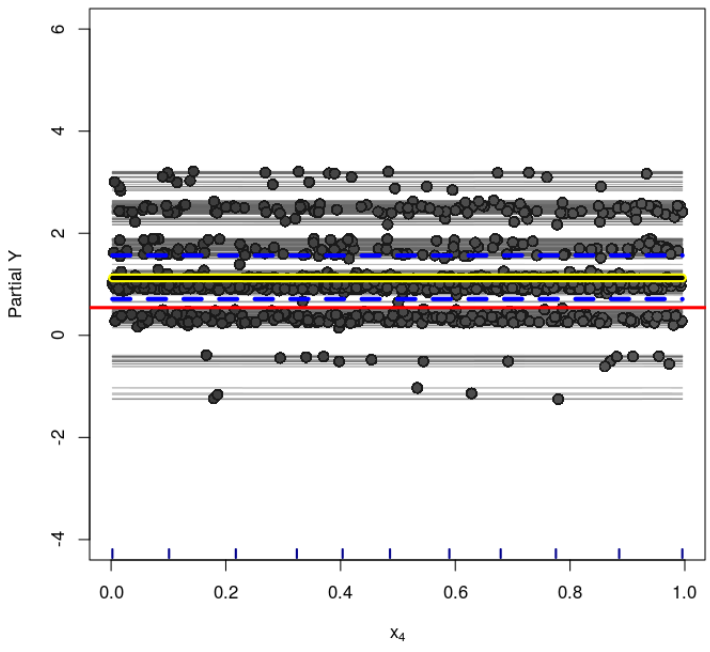

Figure D.75: *PS-BART model (Friedman Function under Sparsity) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

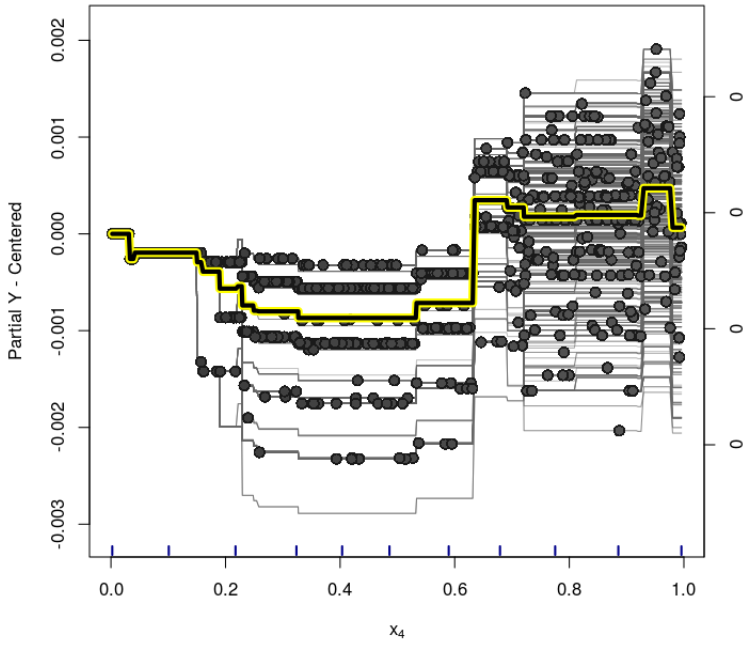

Figure D.76: *PS-BART model (Friedman Function under Sparsity) - Variable x_4 - centered-ICE Plot for the treatment effect.*

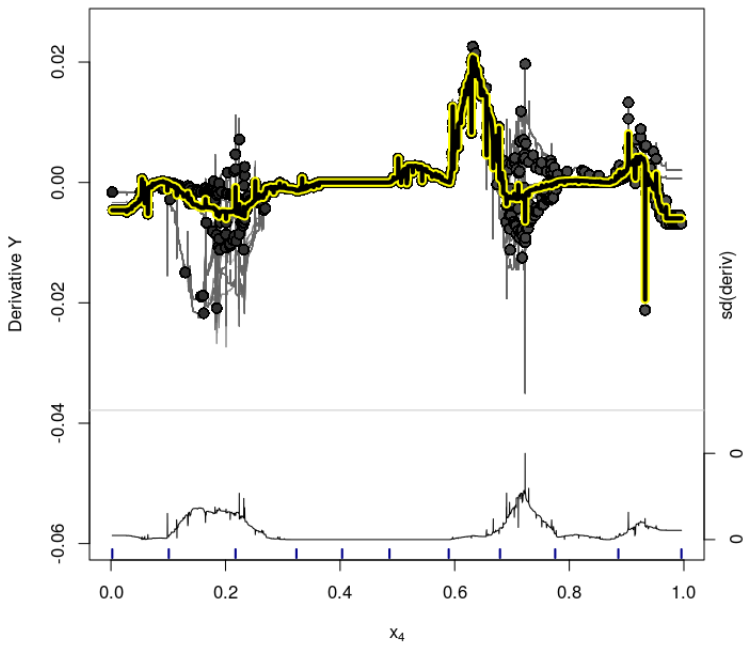

Figure D.77: *PS-BART model (Friedman Function under Sparsity) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

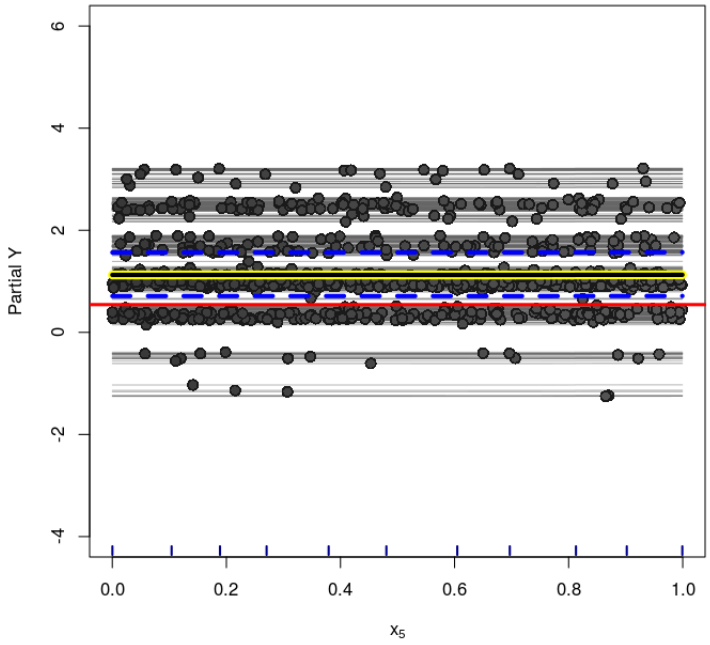

Figure D.78: *PS-BART model (Friedman Function under Sparsity) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

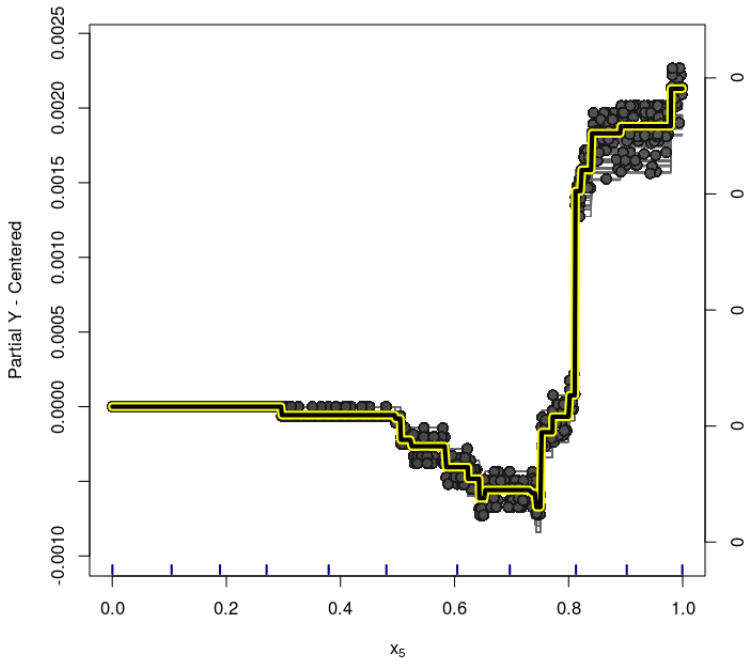

Figure D.79: *PS-BART model (Friedman Function under Sparsity) - Variable x_5 - centered-ICE Plot for the treatment effect.*

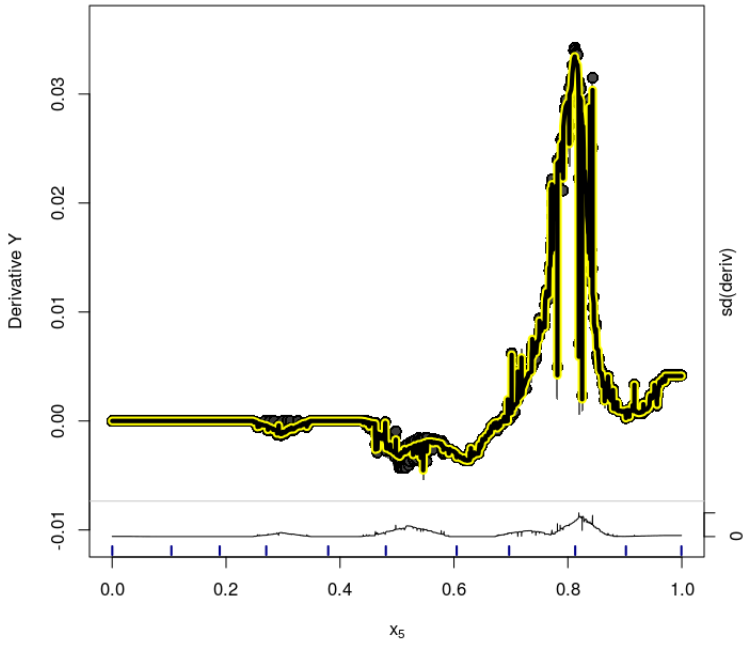

Figure D.80: *PS-BART model (Friedman Function under Sparsity) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

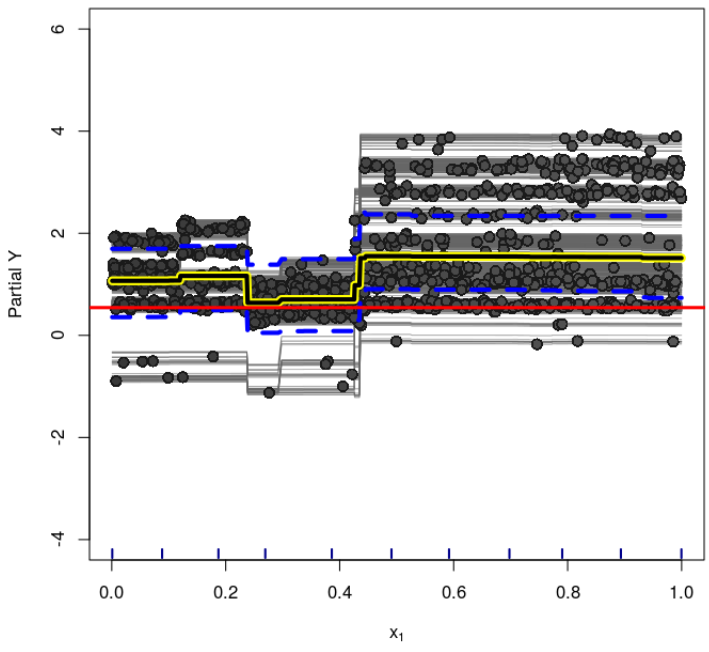

Figure D.81: *GLM-BART model (Friedman Function under Sparsity) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

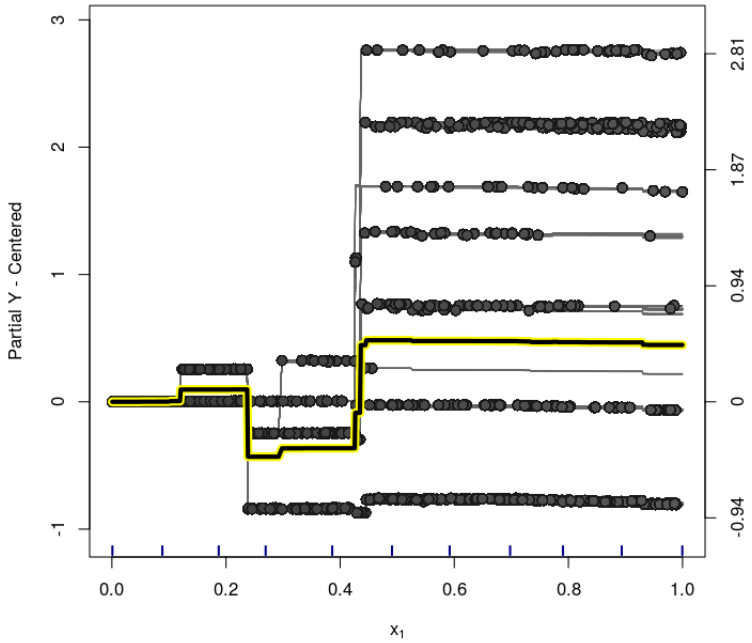

Figure D.82: *GLM-BART model (Friedman Function under Sparsity) - Variable x_1 - centered-ICE Plot for the treatment effect.*

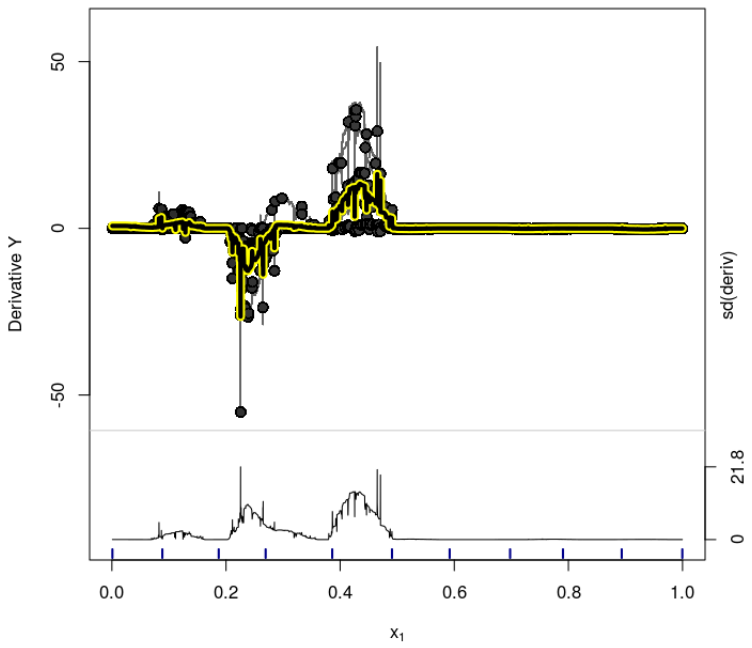

Figure D.83: *GLM-BART model (Friedman Function under Sparsity) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

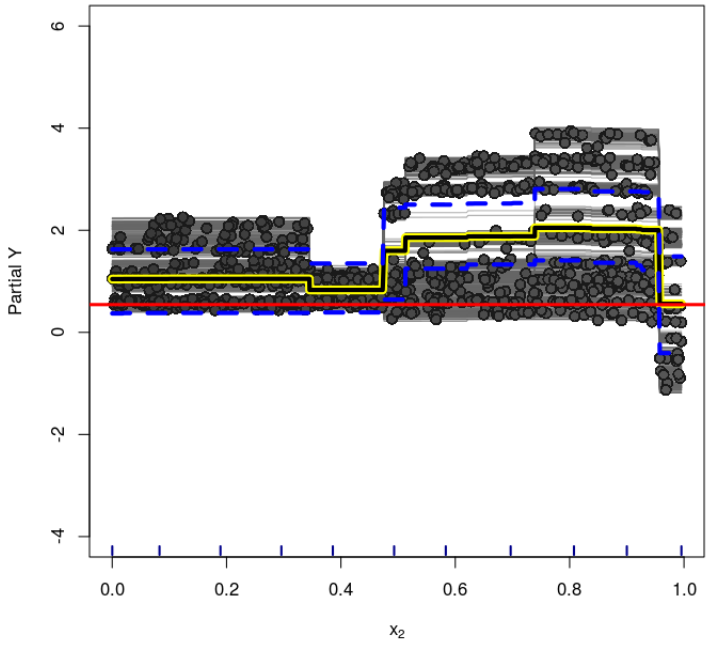

Figure D.84: *GLM-BART model (Friedman Function under Sparsity) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

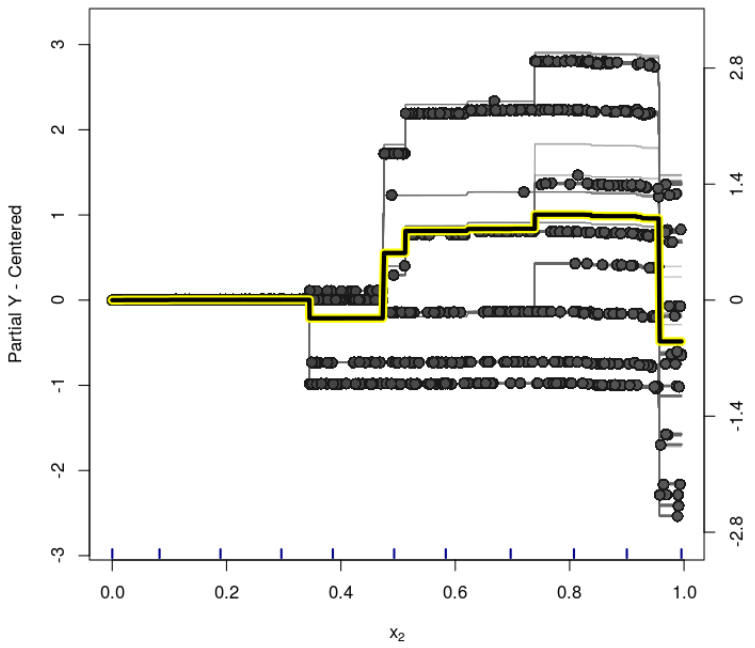

Figure D.85: *GLM-BART model (Friedman Function under Sparsity) - Variable x_2 - centered-ICE Plot for the treatment effect.*

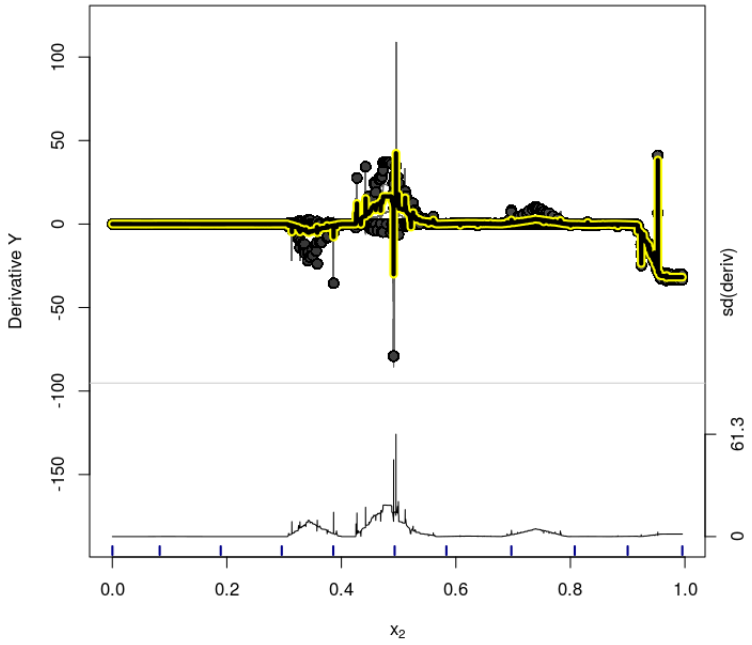

Figure D.86: *GLM-BART model (Friedman Function under Sparsity) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

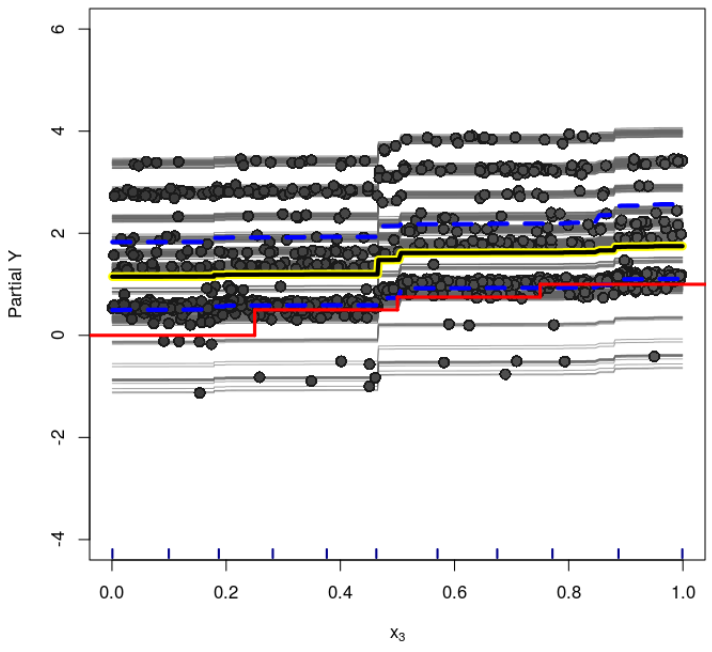

Figure D.87: *GLM-BART model (Friedman Function under Sparsity) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

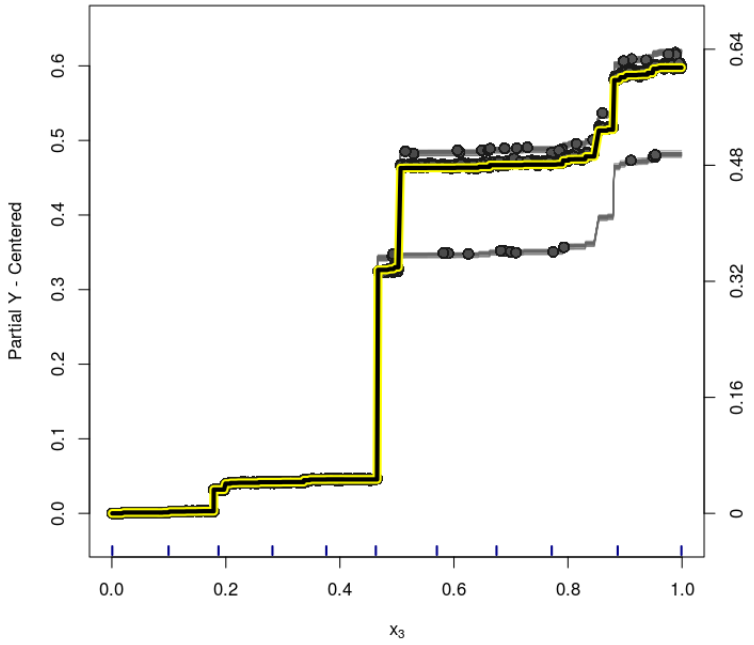

Figure D.88: *GLM-BART model (Friedman Function under Sparsity) - Variable x_3 - centered-ICE Plot for the treatment effect.*

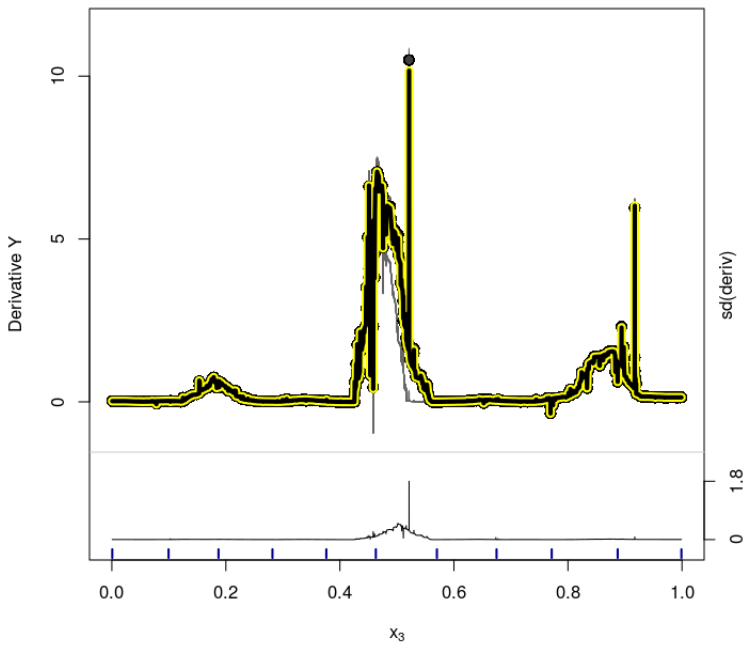

Figure D.89: *GLM-BART model (Friedman Function under Sparsity) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

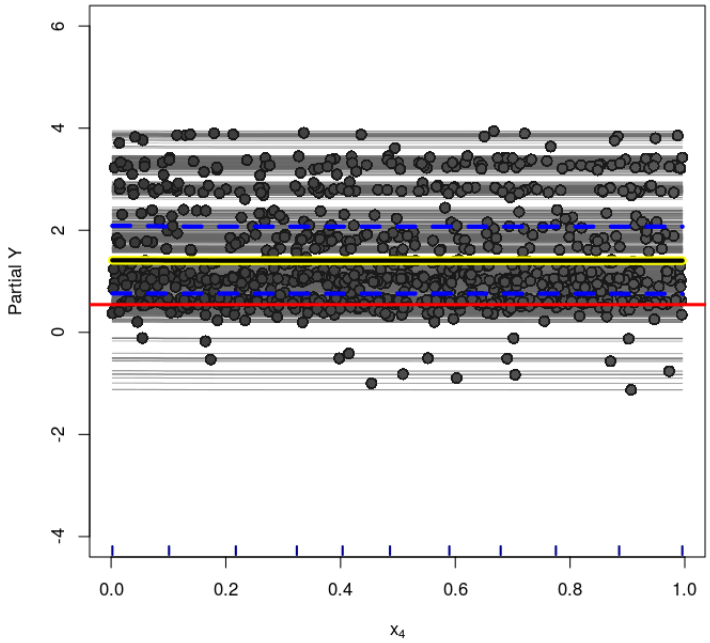

Figure D.90: *GLM-BART model (Friedman Function under Sparsity) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

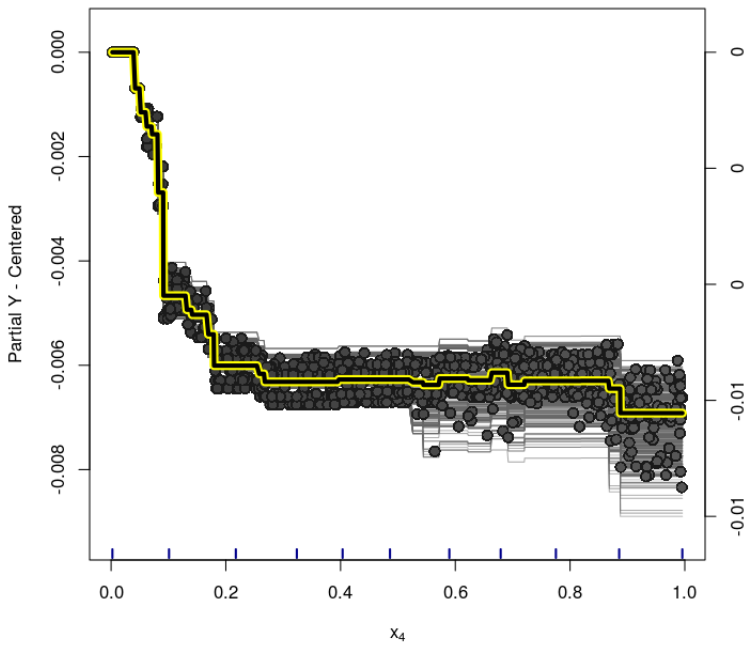

Figure D.91: *GLM-BART model (Friedman Function under Sparsity) - Variable x_4 - centered-ICE Plot for the treatment effect.*

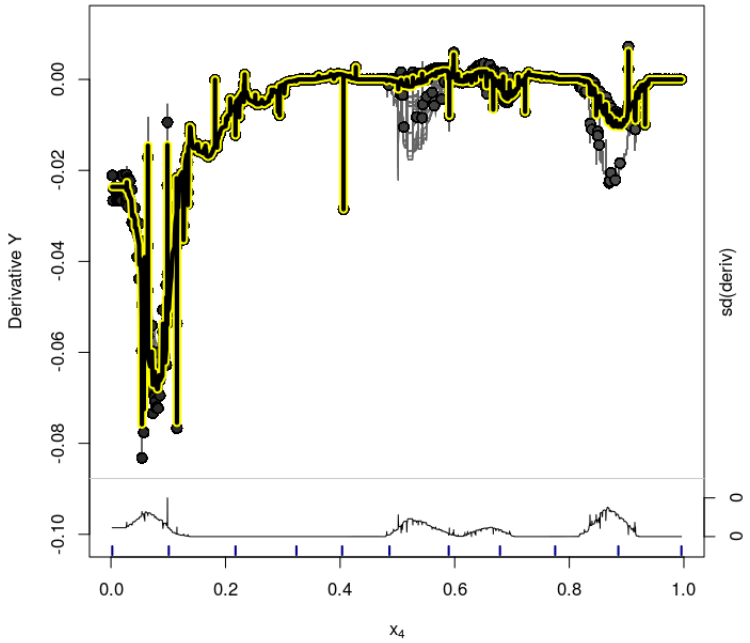

Figure D.92: *GLM-BART model (Friedman Function under Sparsity) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

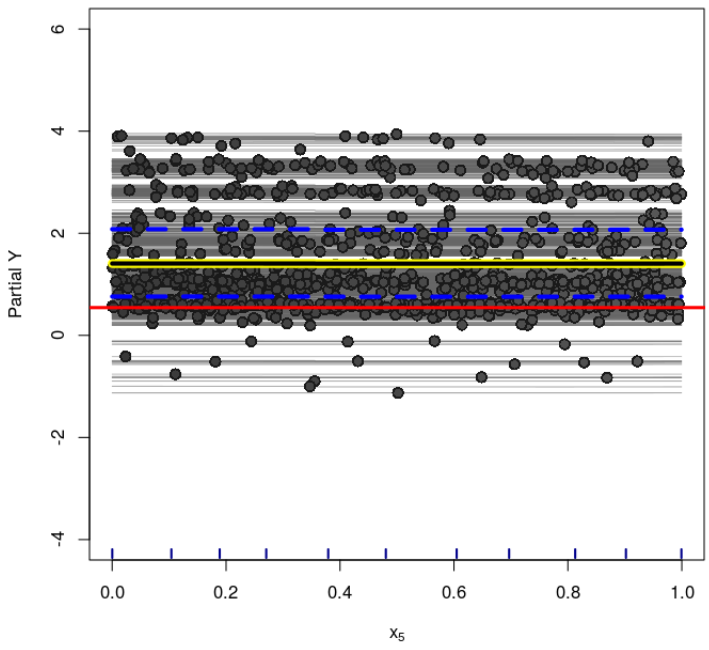

Figure D.93: *GLM-BART model (Friedman Function under Sparsity) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

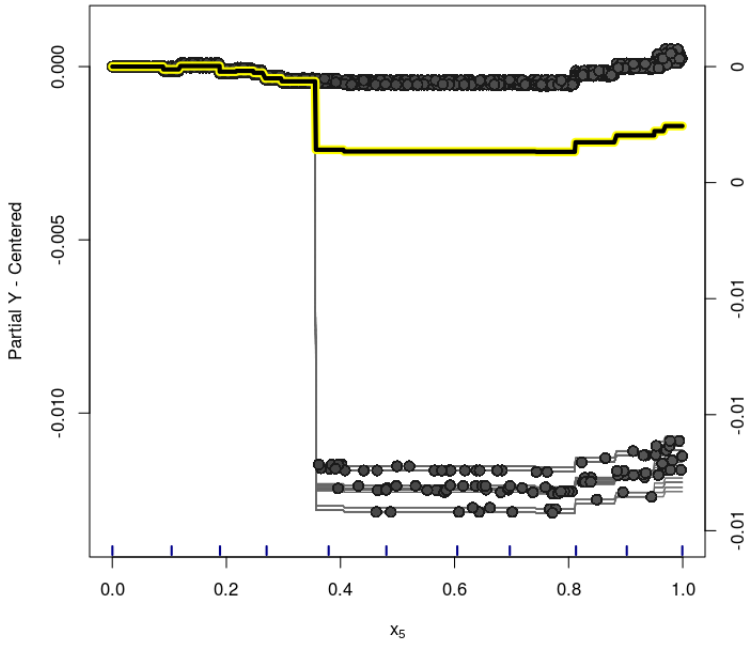

Figure D.94: *GLM-BART model (Friedman Function under Sparsity) - Variable x_5 - centered-ICE Plot for the treatment effect.*

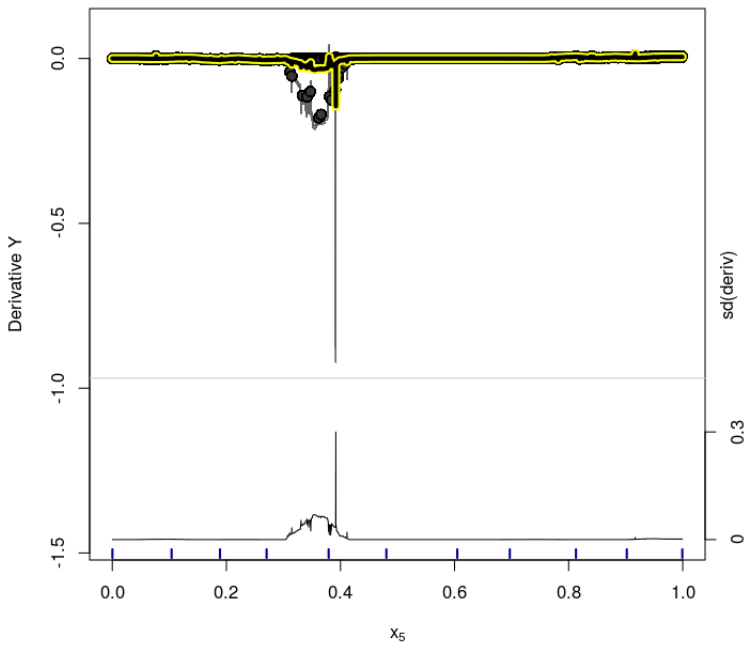

Figure D.95: *GLM-BART model (Friedman Function under Sparsity) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

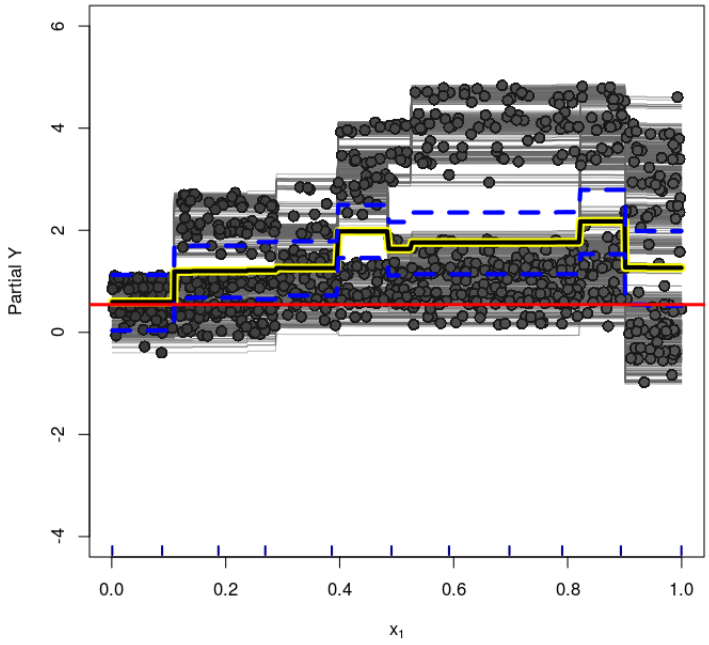

Figure D.96: *Rand-BART model (Friedman Function under Sparsity) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

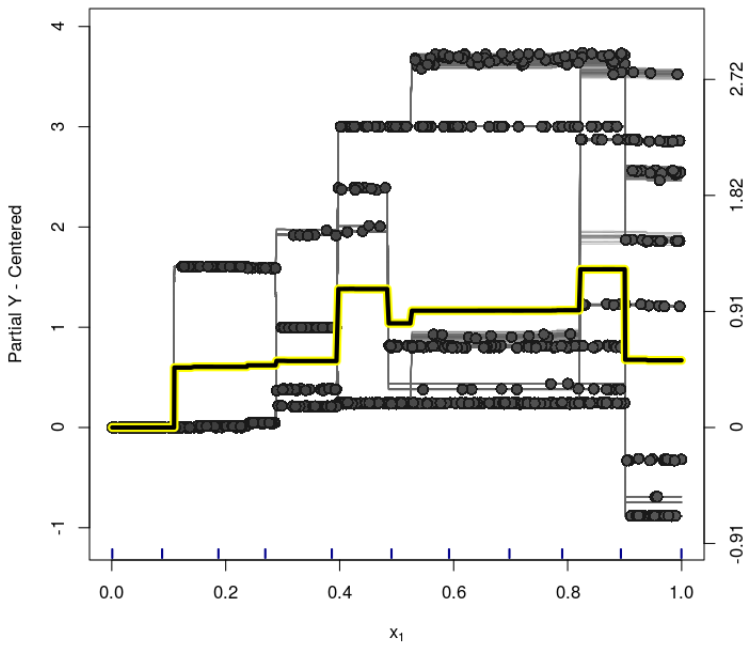

Figure D.97: *Rand-BART model (Friedman Function under Sparsity) - Variable x_1 - centered-ICE Plot for the treatment effect.*

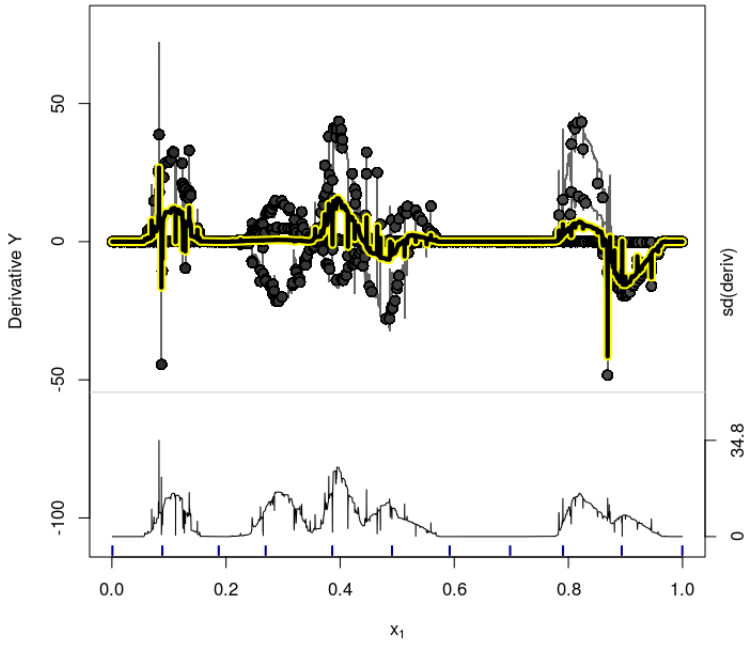

Figure D.98: *Rand-BART model (Friedman Function under Sparsity) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

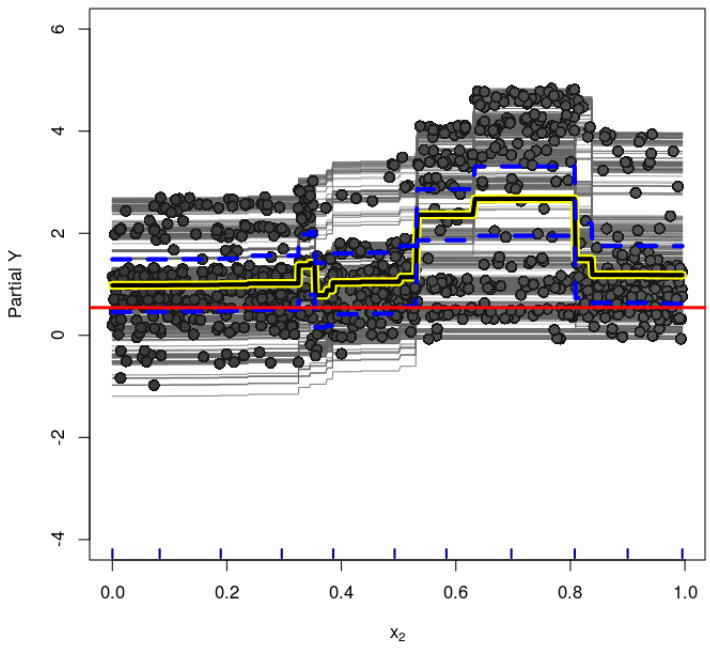

Figure D.99: *Rand-BART model (Friedman Function under Sparsity) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

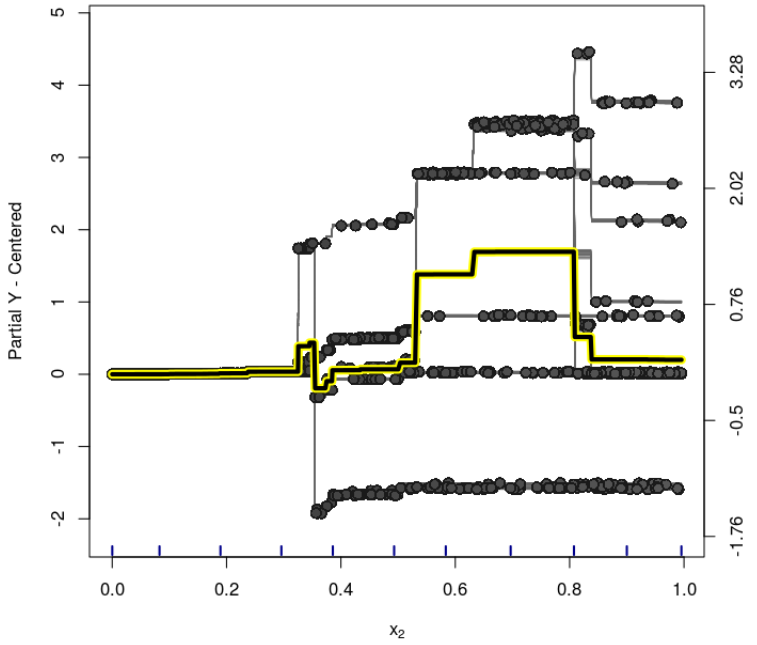

Figure D.100: *Rand-BART model (Friedman Function under Sparsity) - Variable x_2 - centered-ICE Plot for the treatment effect.*

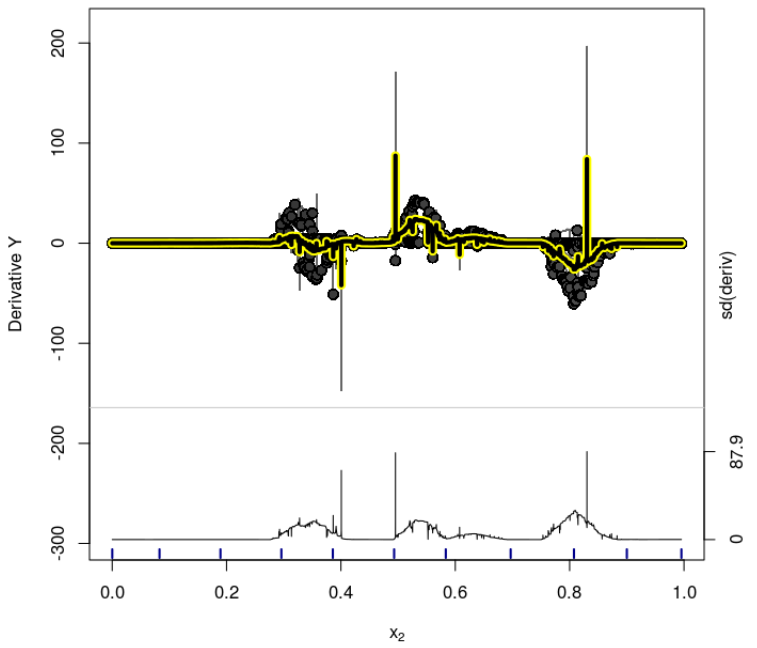

Figure D.101: *Rand-BART model (Friedman Function under Sparsity) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

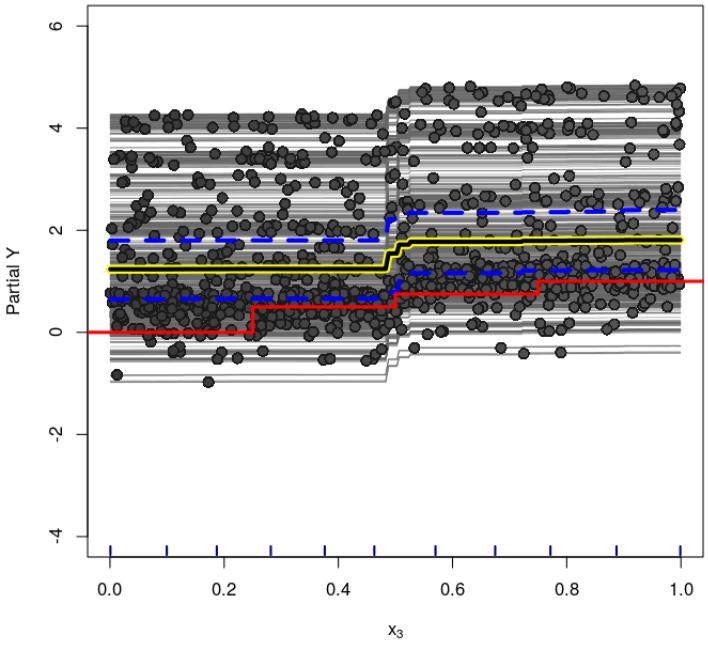

Figure D.102: *Rand-BART model (Friedman Function under Sparsity) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

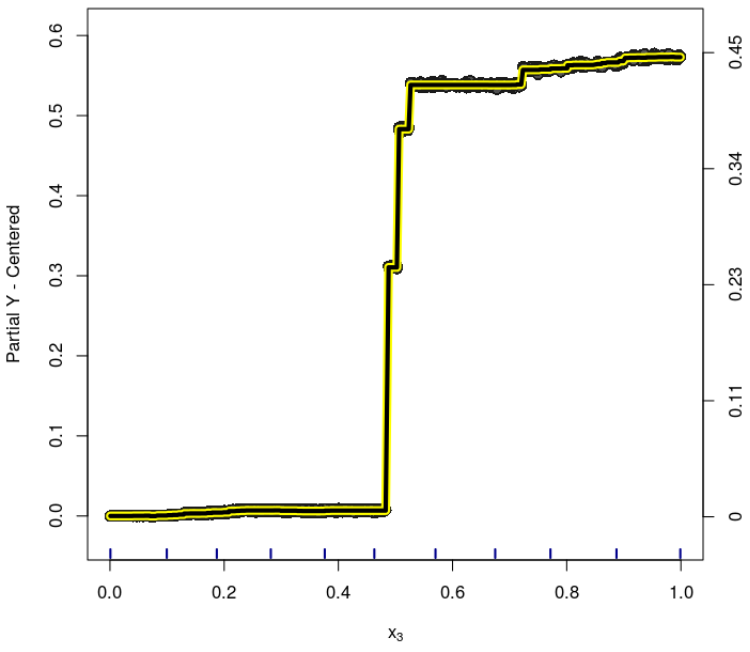

Figure D.103: *Rand-BART model (Friedman Function under Sparsity) - Variable x_3 - centered-ICE Plot for the treatment effect.*

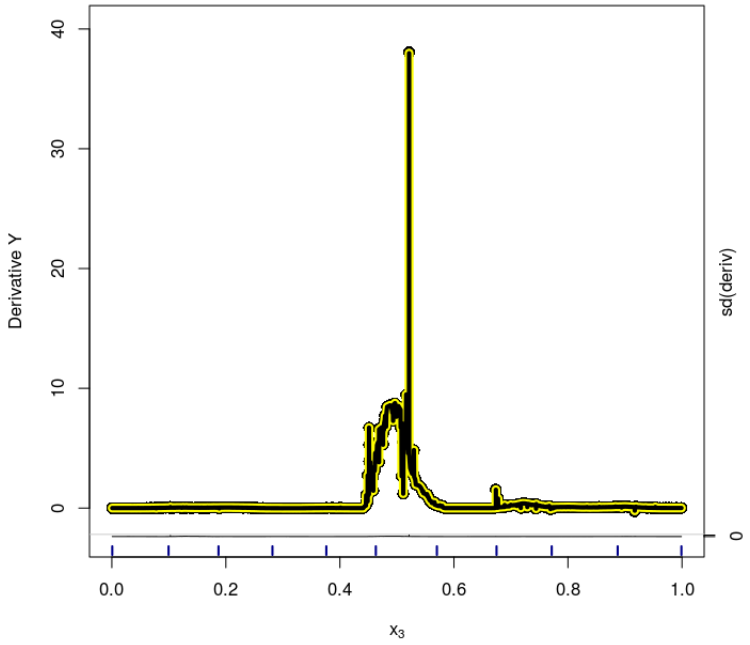

Figure D.104: *Rand-BART model (Friedman Function under Sparsity) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

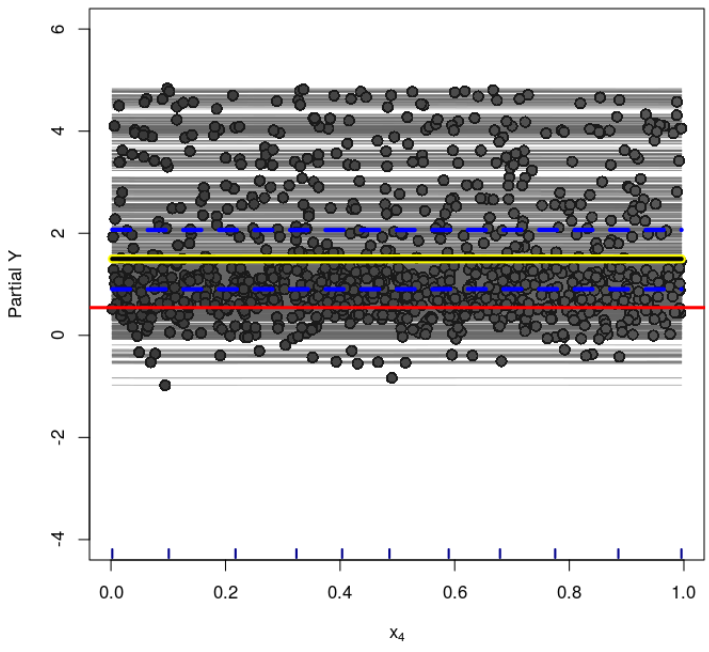

Figure D.105: *Rand-BART model (Friedman Function under Sparsity) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

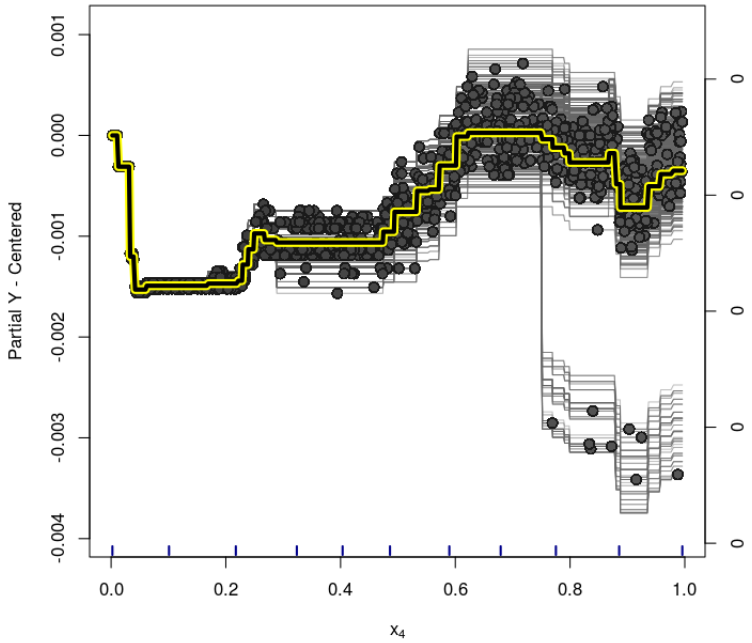

Figure D.106: *Rand-BART model (Friedman Function under Sparsity) - Variable x_4 - centered-ICE Plot for the treatment effect.*

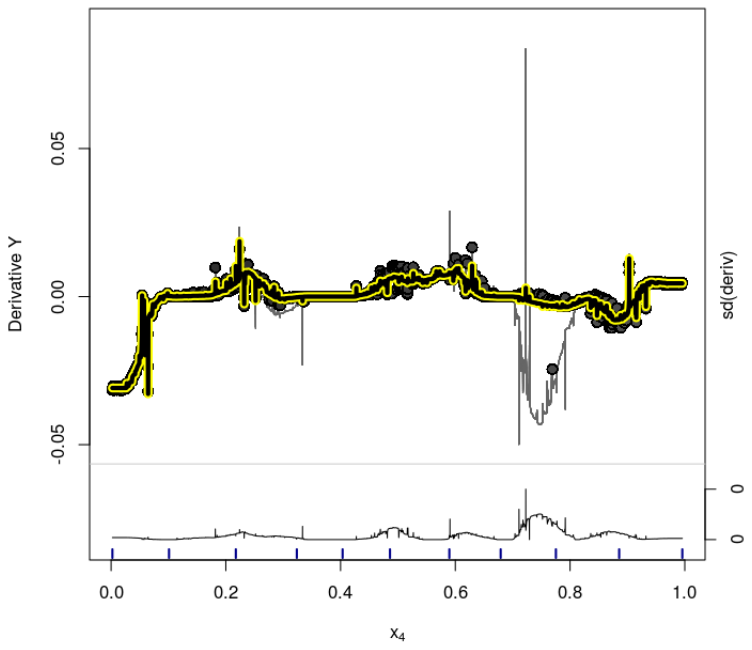

Figure D.107: *Rand-BART model (Friedman Function under Sparsity) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

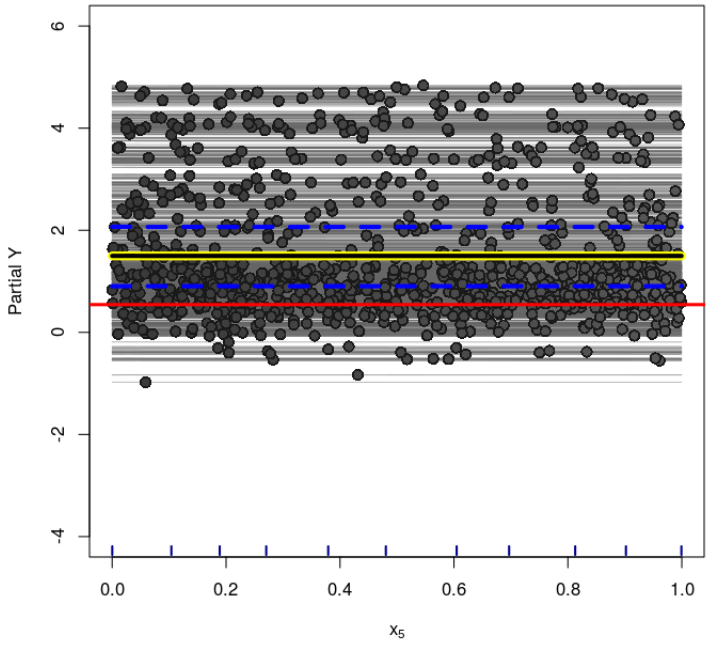

Figure D.108: *Rand-BART model (Friedman Function under Sparsity) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

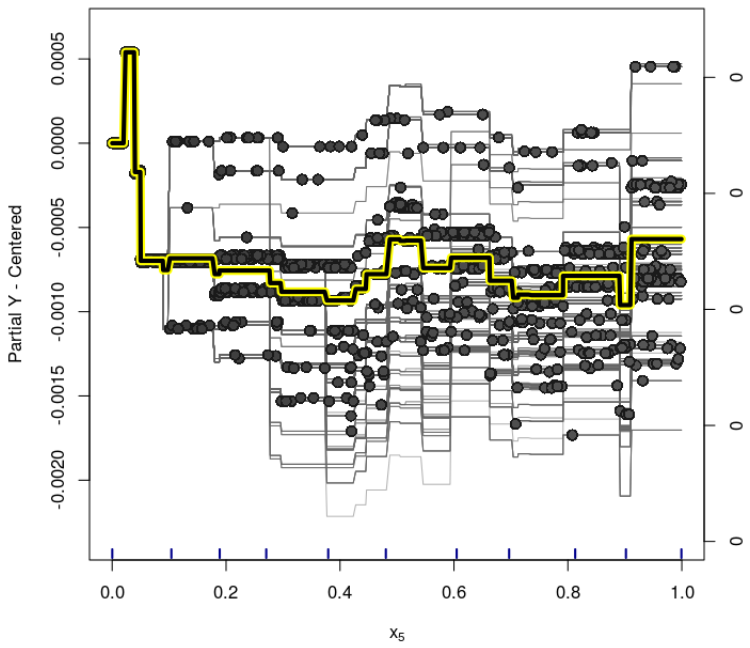

Figure D.109: *Rand-BART model (Friedman Function under Sparsity) - Variable x_5 - centered-ICE Plot for the treatment effect.*

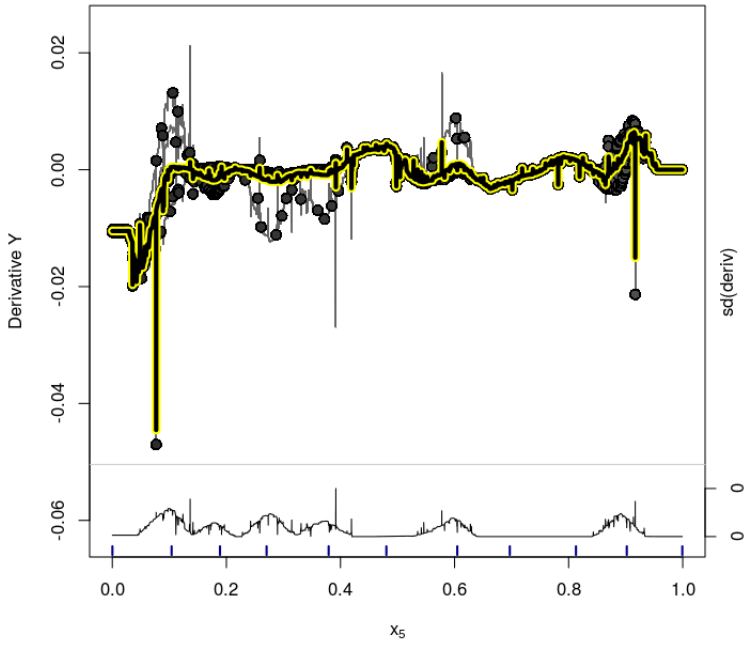

Figure D.110: *Rand-BART model (Friedman Function under Sparsity) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

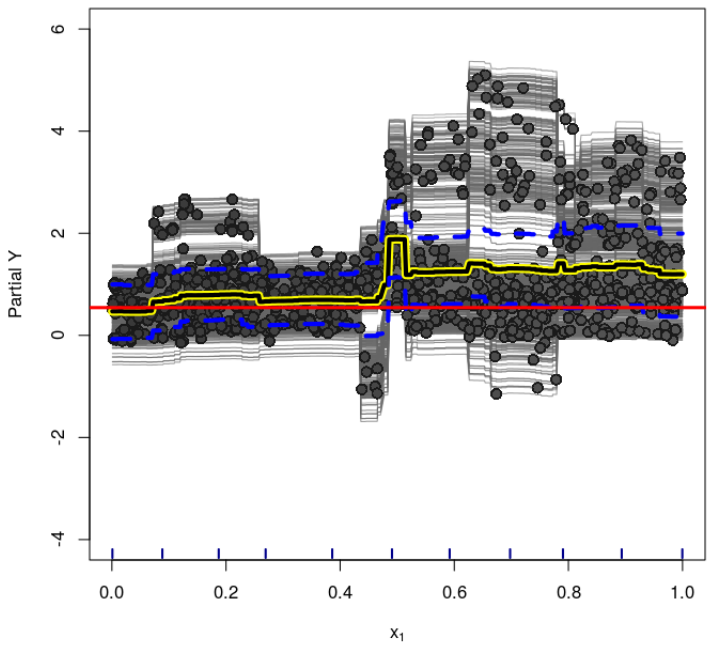

Figure D.111: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

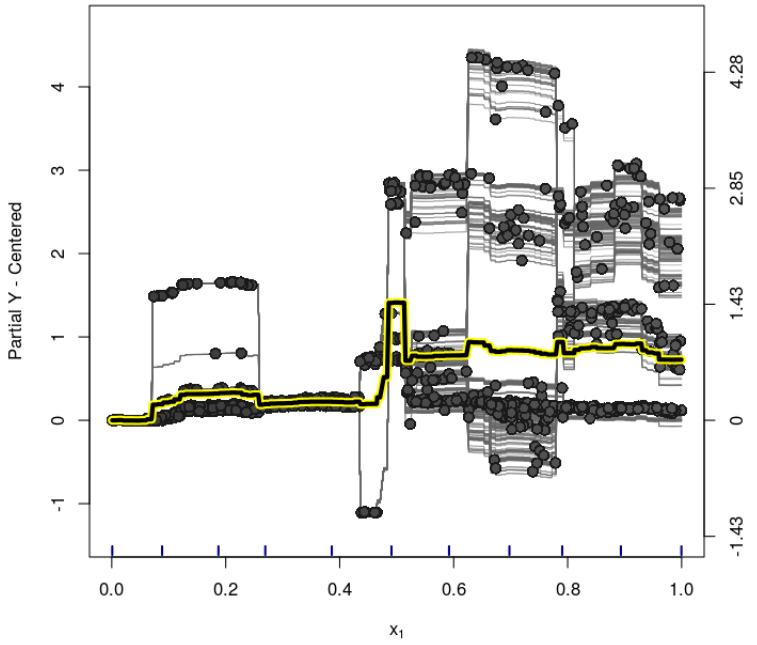

Figure D.112: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_1 - centered-ICE Plot for the treatment effect.*

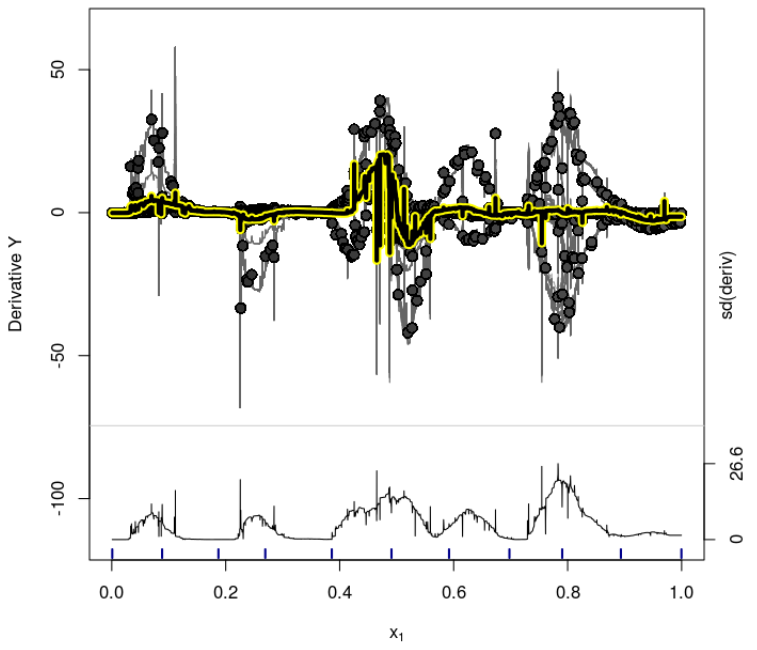

Figure D.113: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

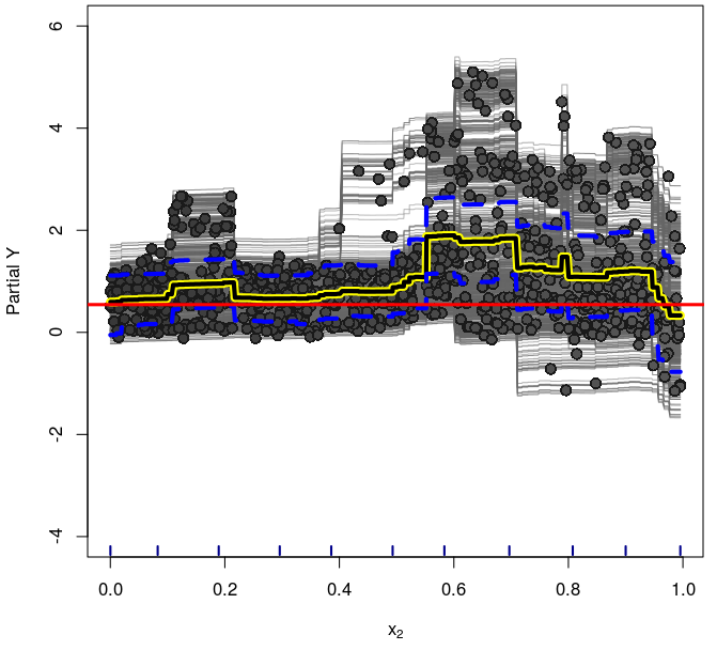

Figure D.114: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

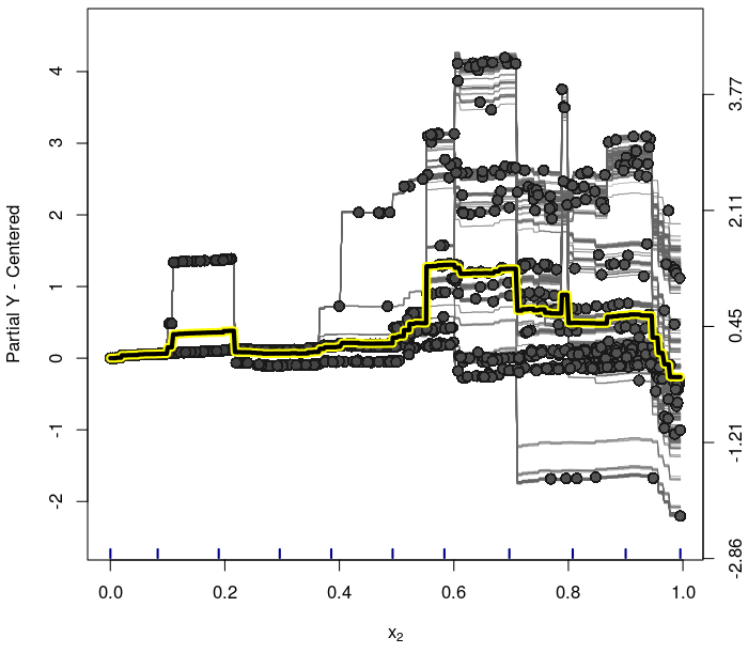

Figure D.115: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_2 - centered-ICE Plot for the treatment effect.*

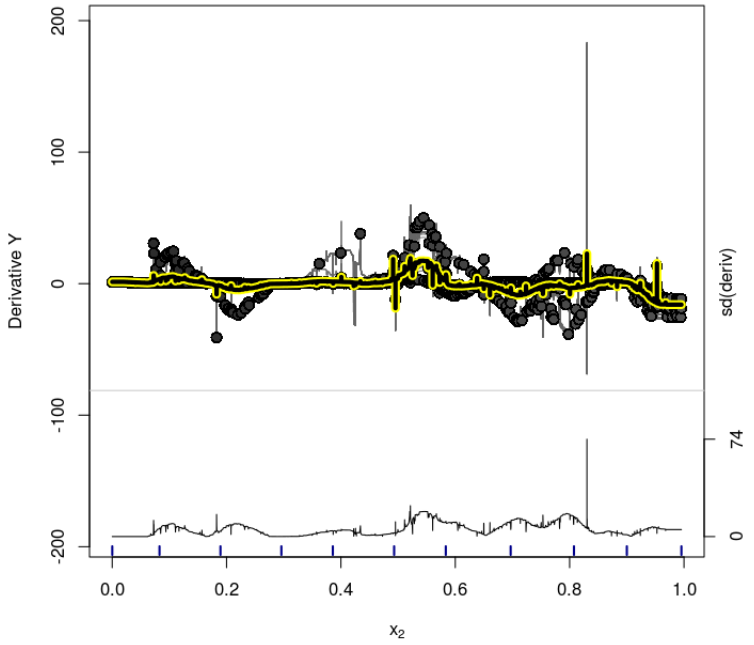

Figure D.116: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

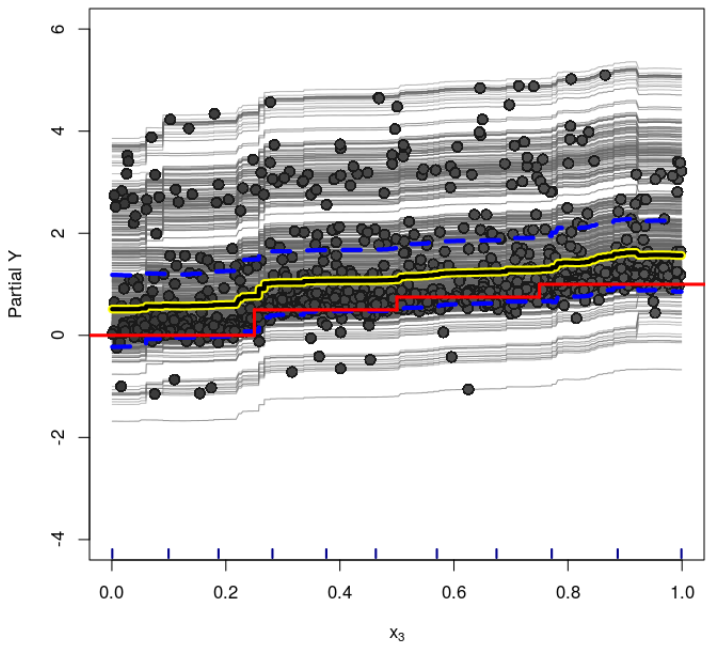

Figure D.117: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

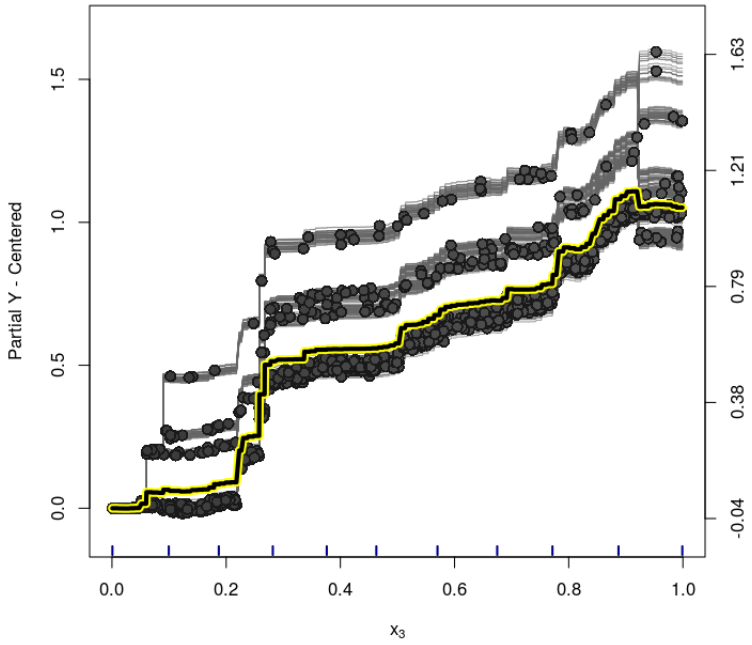

Figure D.118: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_3 - centered-ICE Plot for the treatment effect.*

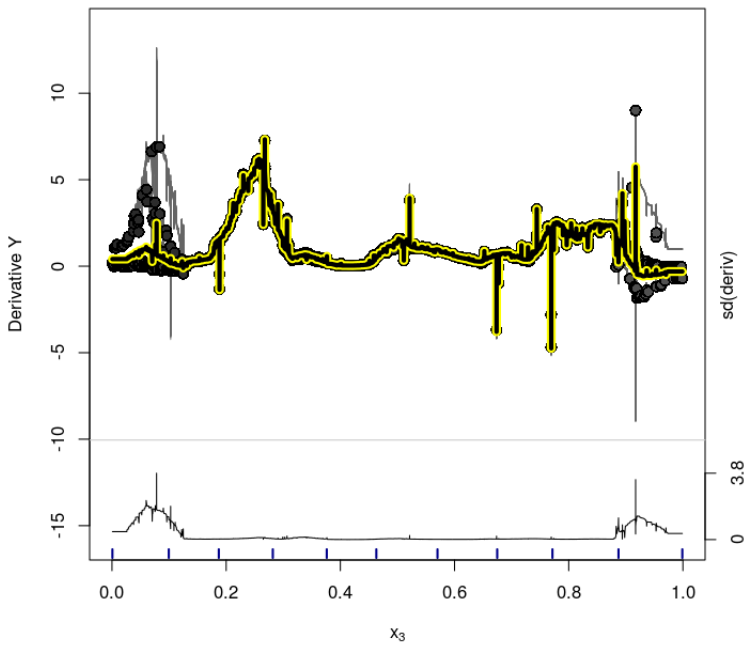

Figure D.119: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

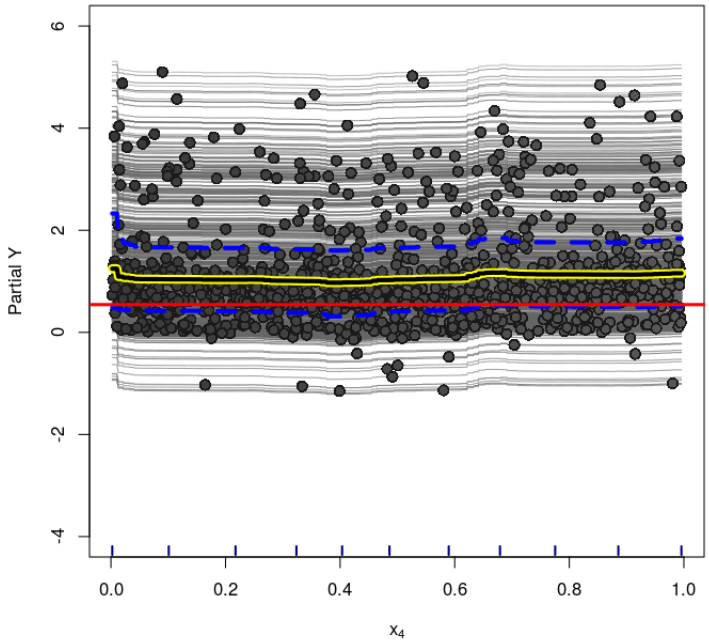

Figure D.120: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

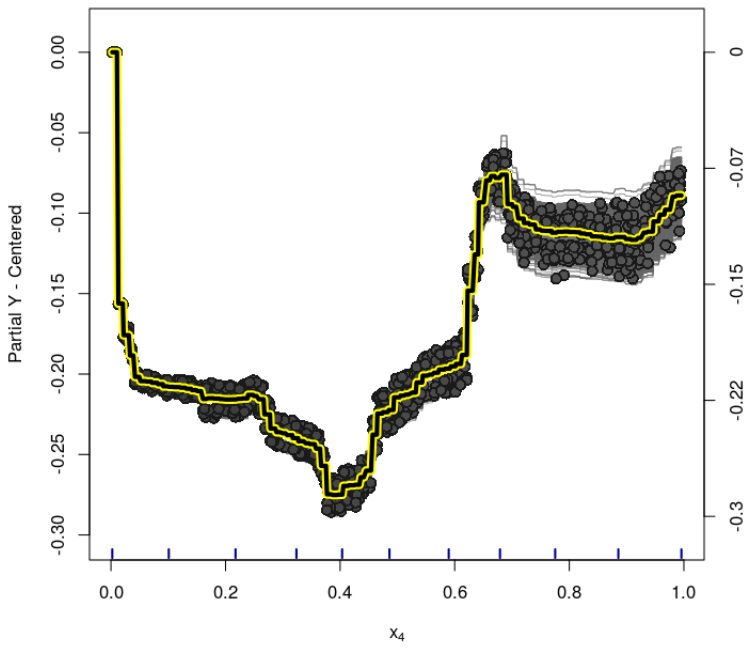

Figure D.121: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_4 - centered-ICE Plot for the treatment effect.*

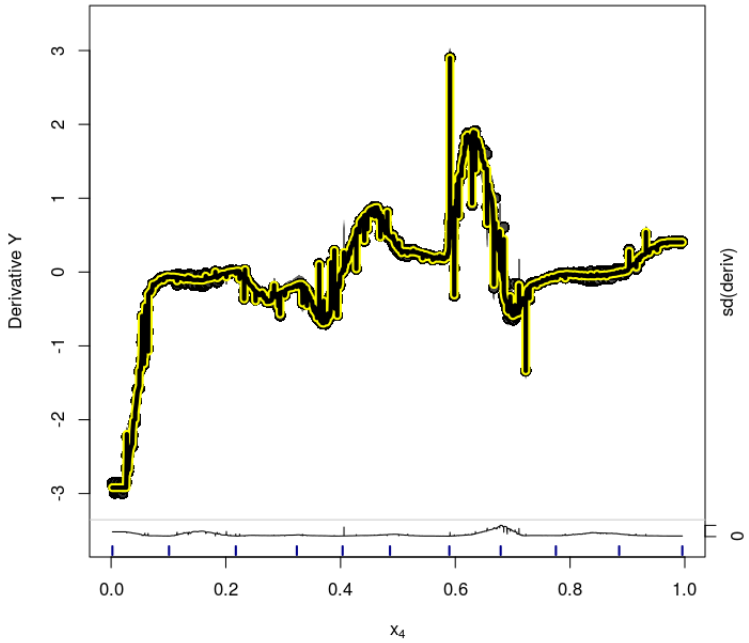

Figure D.122: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

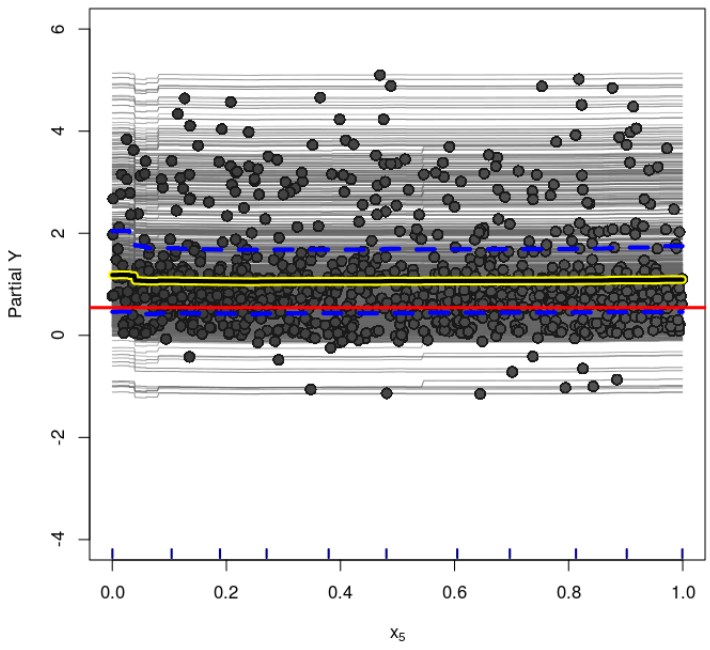

Figure D.123: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

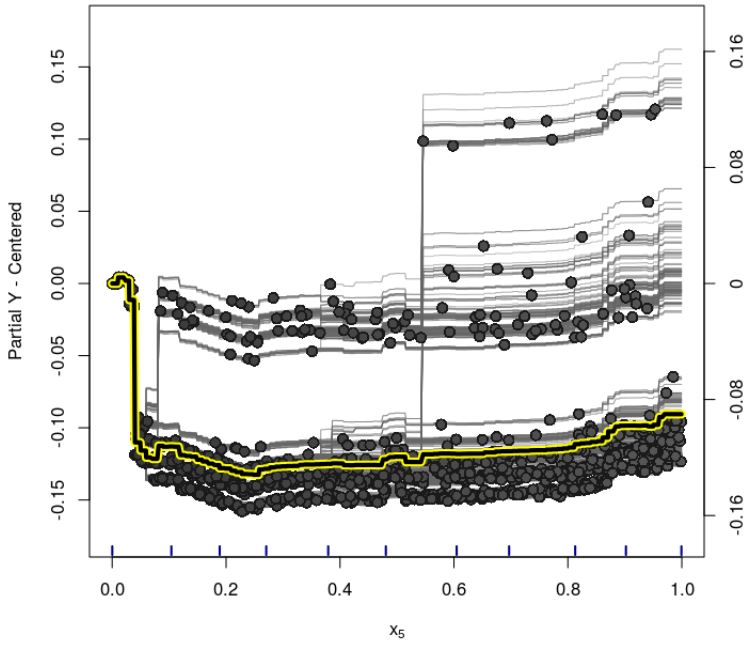

Figure D.124: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_5 - centered-ICE Plot for the treatment effect.*

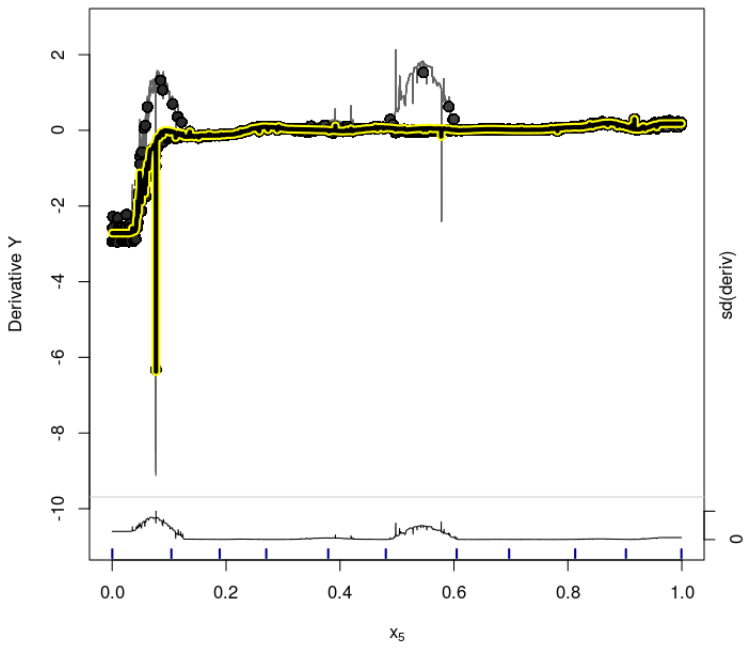

Figure D.125: *Vanilla-DART model (Friedman Function under Sparsity) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

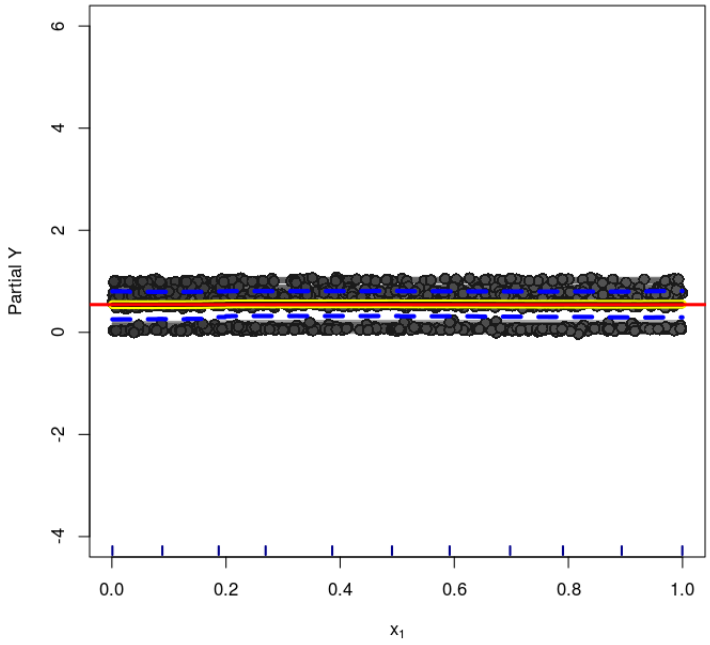

Figure D.126: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

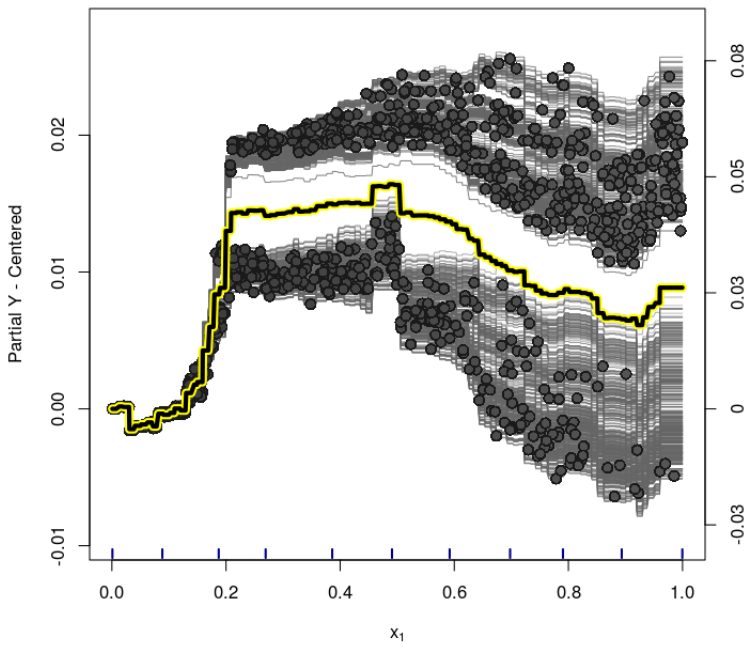

Figure D.127: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_1 - centered-ICE Plot for the treatment effect.*

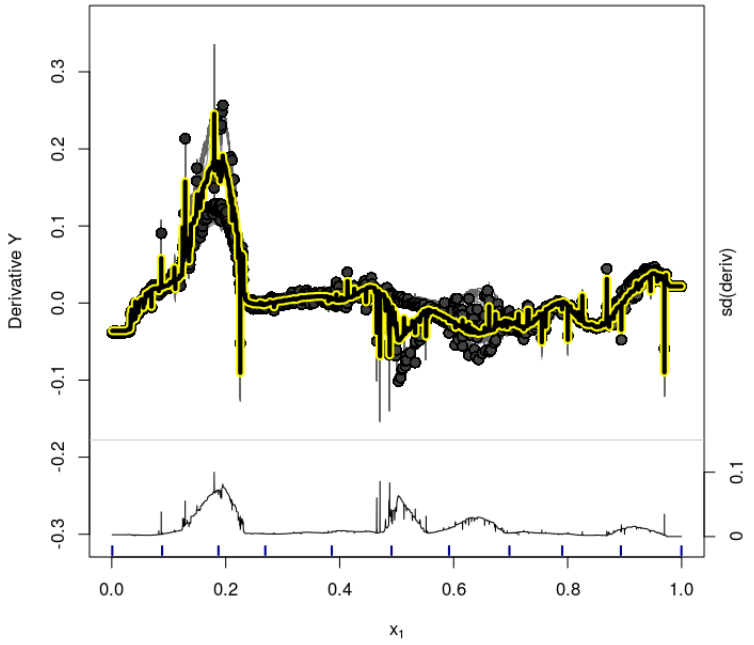

Figure D.128: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

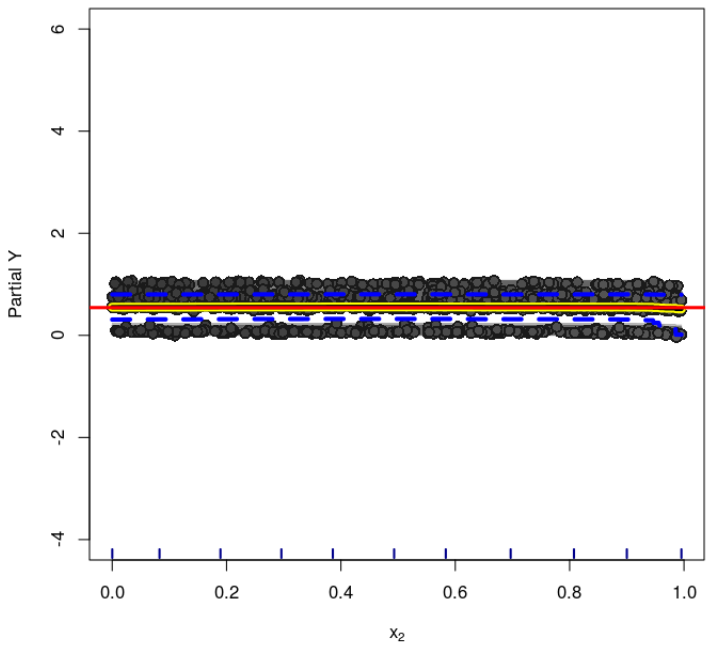

Figure D.129: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

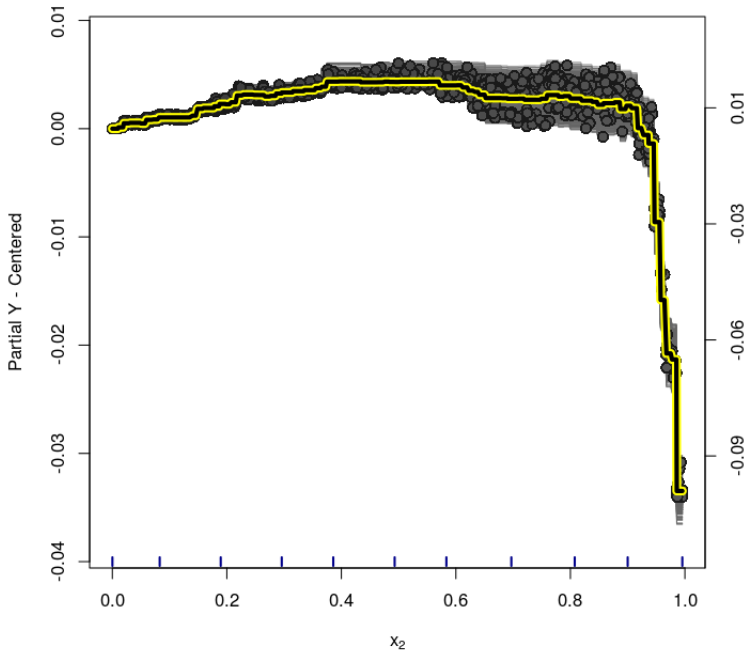

Figure D.130: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_2 - centered-ICE Plot for the treatment effect.*

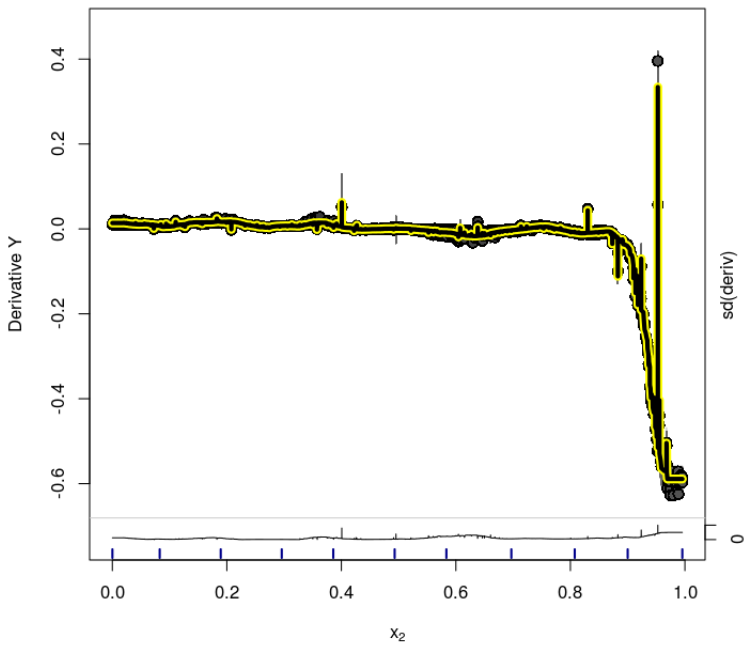

Figure D.131: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

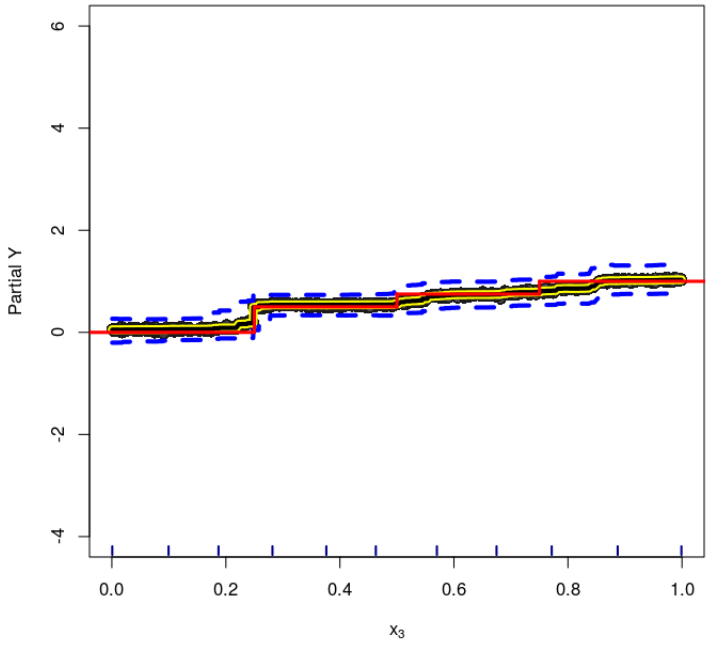

Figure D.132: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

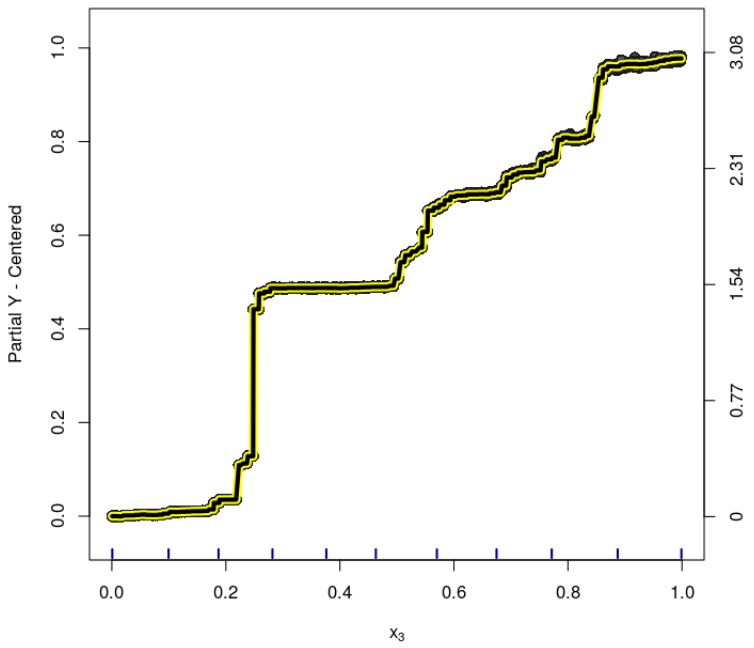

Figure D.133: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_3 - centered-ICE Plot for the treatment effect.*

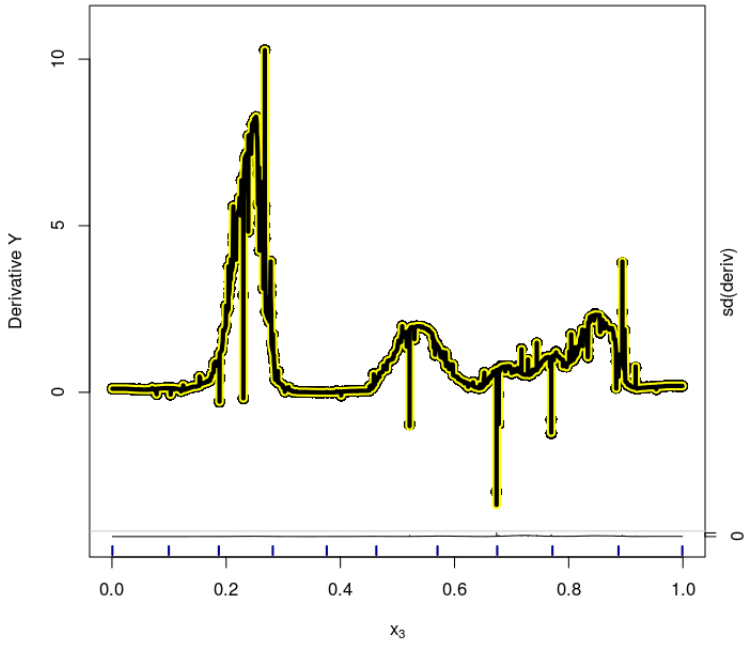

Figure D.134: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

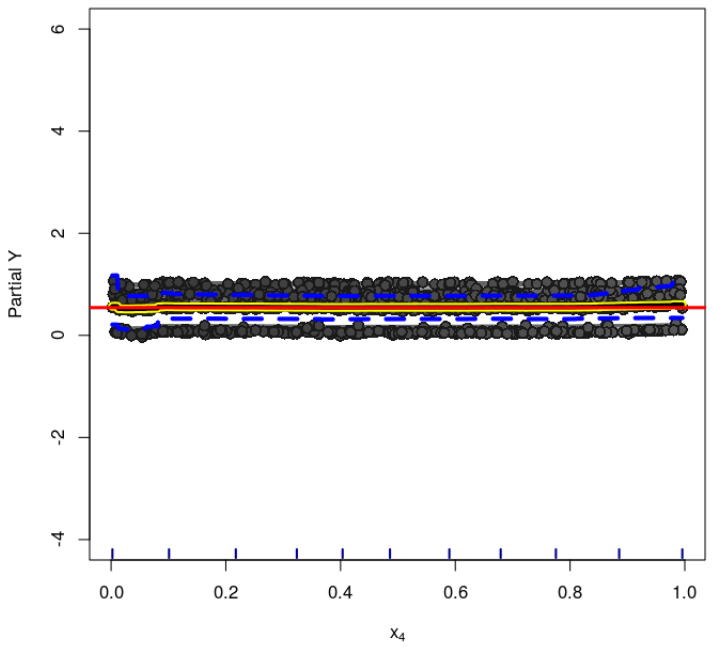

Figure D.135: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

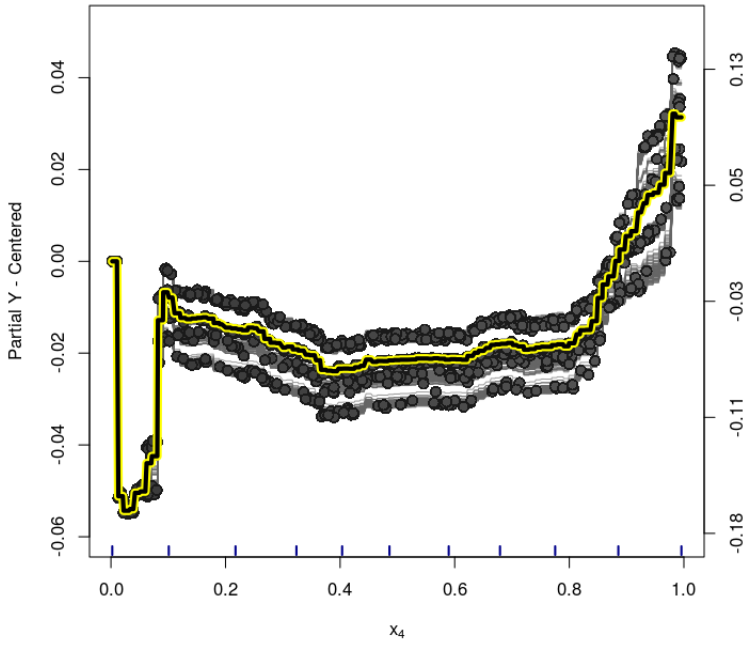

Figure D.136: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_4 - centered-ICE Plot for the treatment effect.*

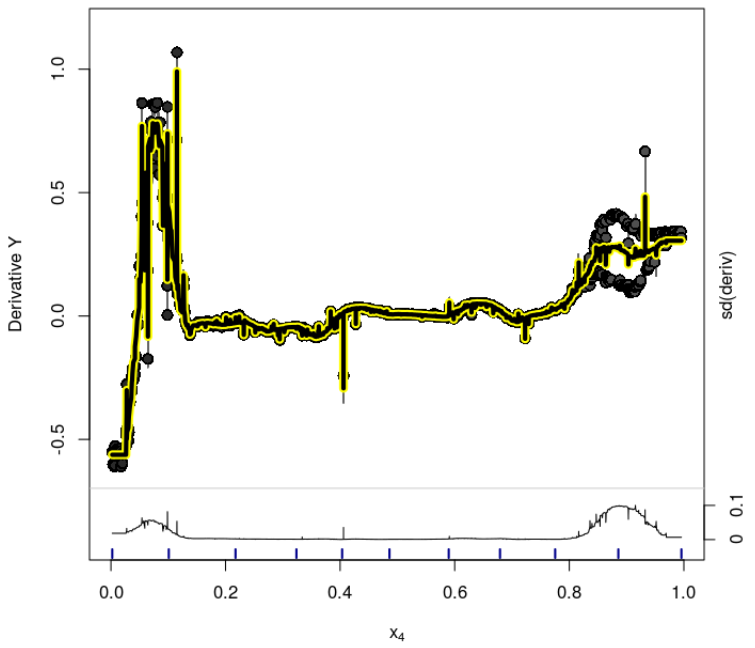

Figure D.137: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

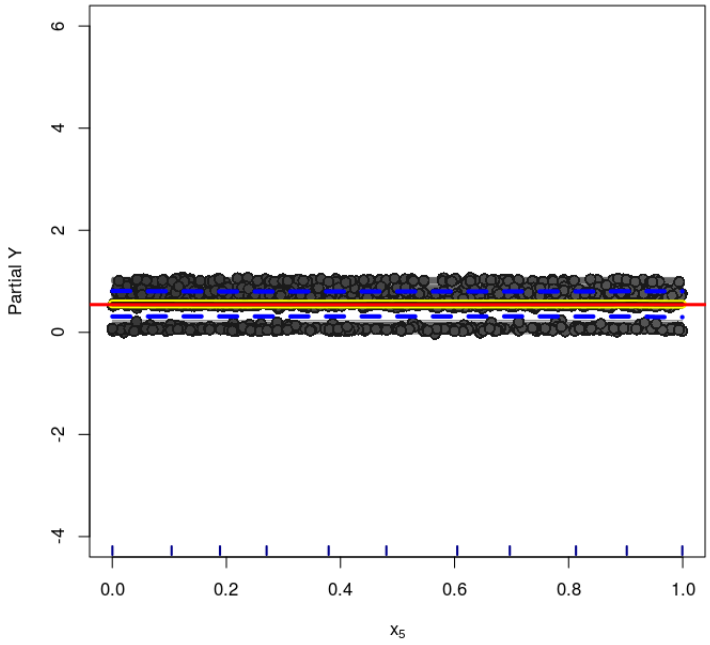

Figure D.138: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

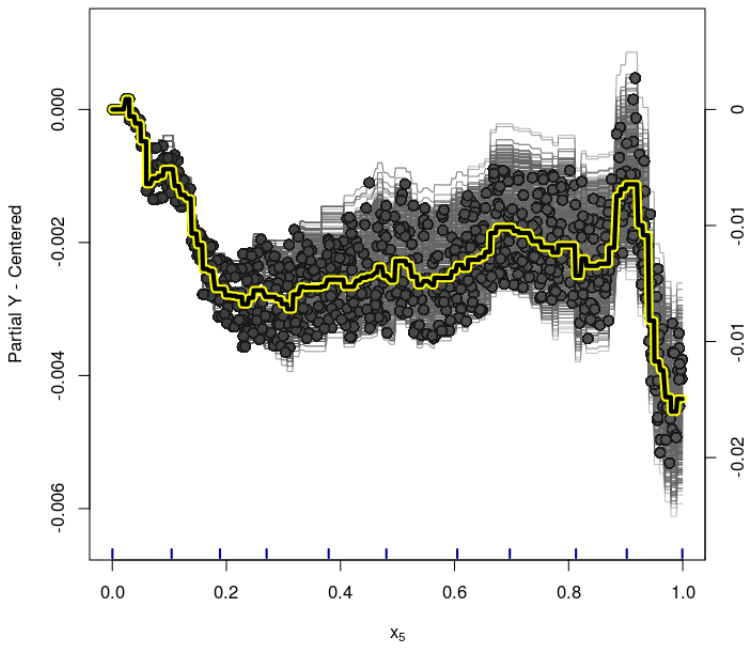

Figure D.139: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_5 - centered-ICE Plot for the treatment effect.*

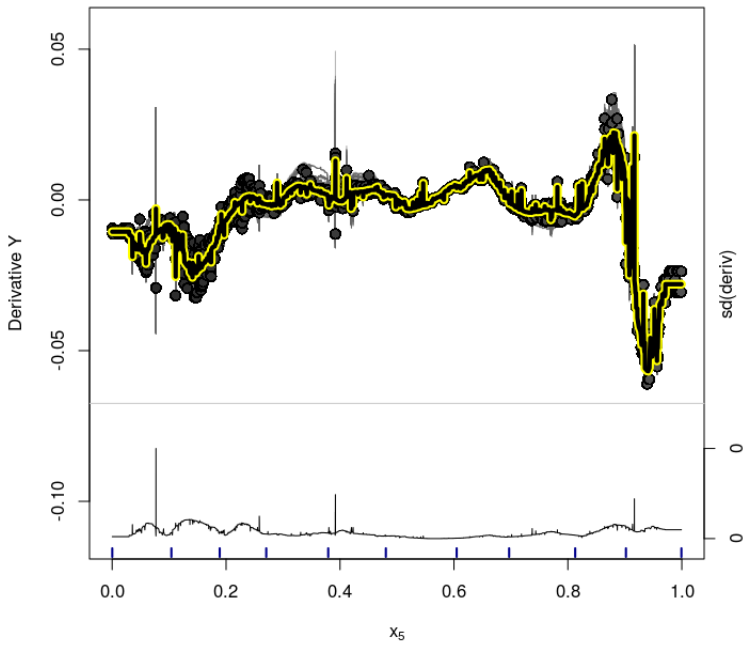

Figure D.140: *Oracle-DART model (Friedman Function under Sparsity) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

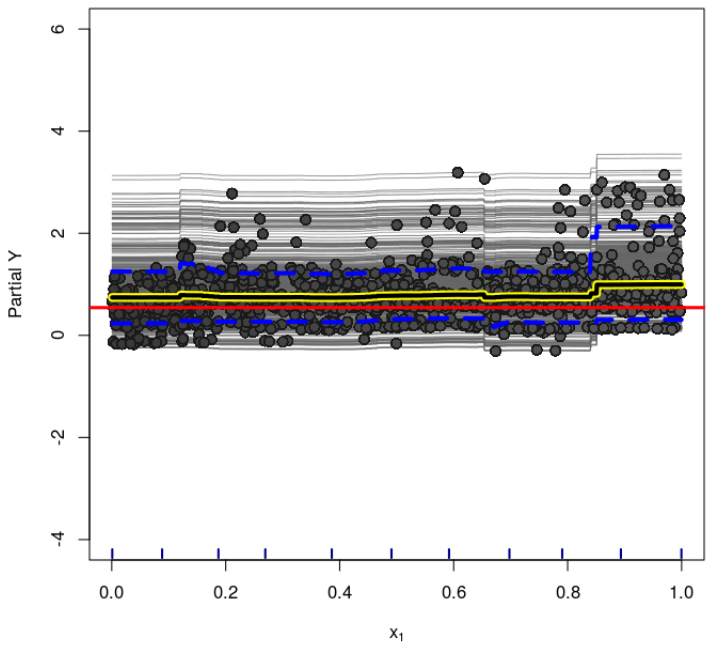

Figure D.141: *PS-DART model (Friedman Function under Sparsity) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

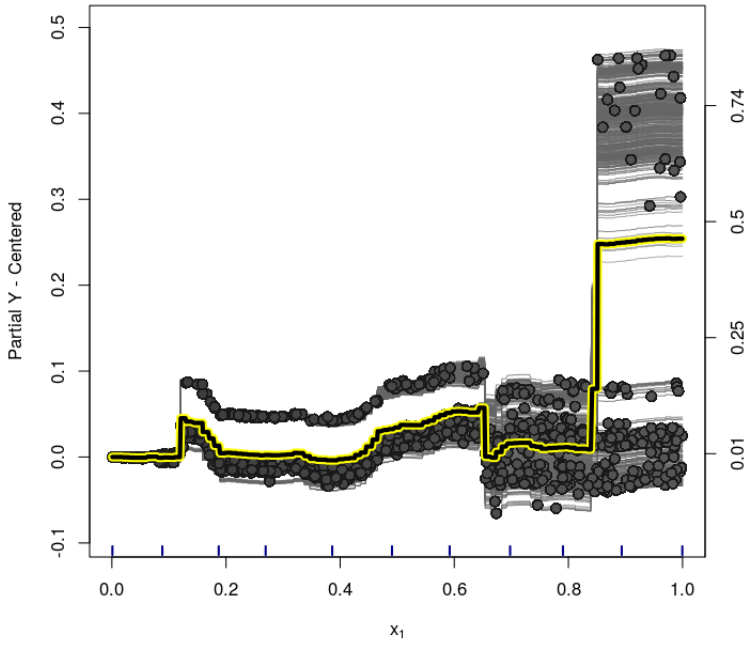

Figure D.142: *PS-DART model (Friedman Function under Sparsity) - Variable x_1 - centered-ICE Plot for the treatment effect.*

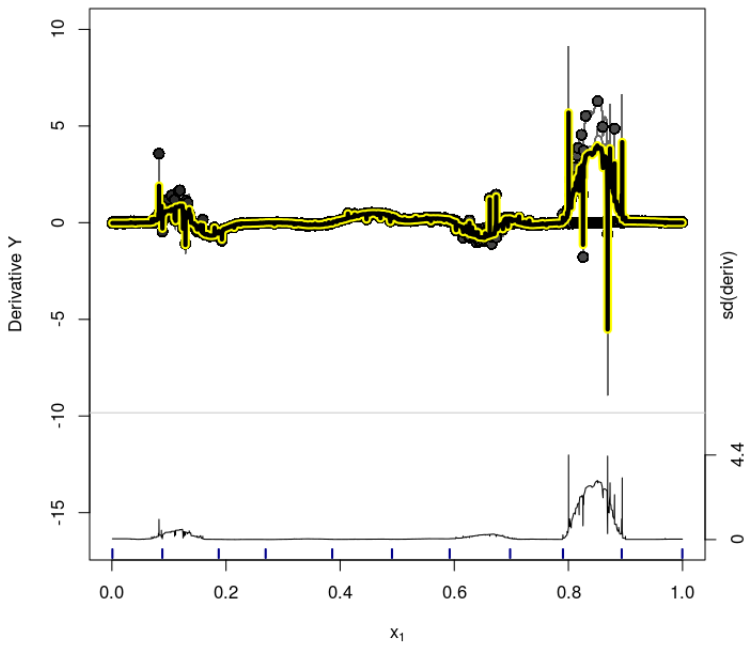

Figure D.143: *PS-DART model (Friedman Function under Sparsity) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

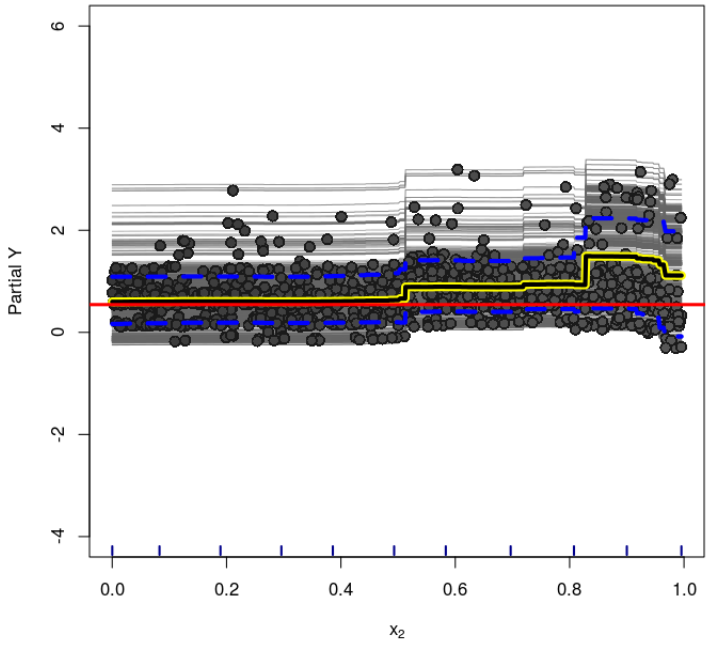

Figure D.144: *PS-DART model (Friedman Function under Sparsity) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

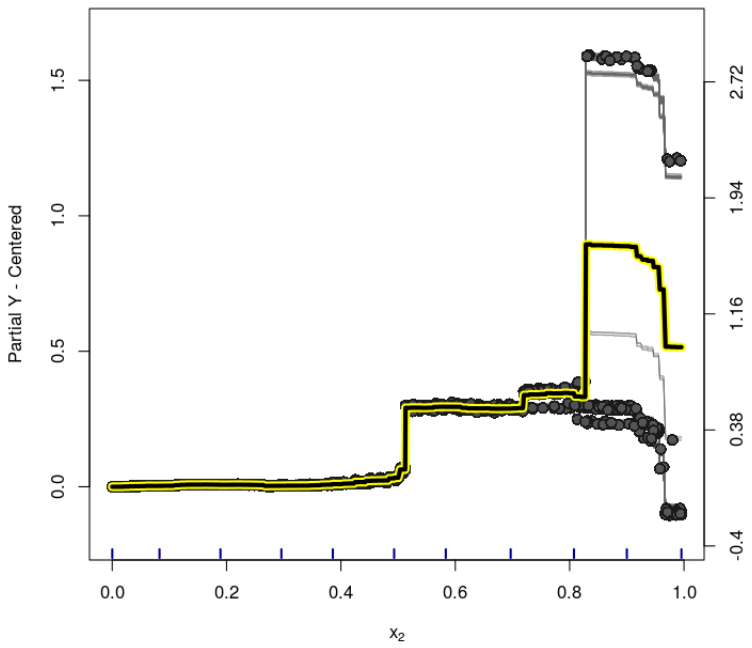

Figure D.145: *PS-DART model (Friedman Function under Sparsity) - Variable x_2 - centered-ICE Plot for the treatment effect.*

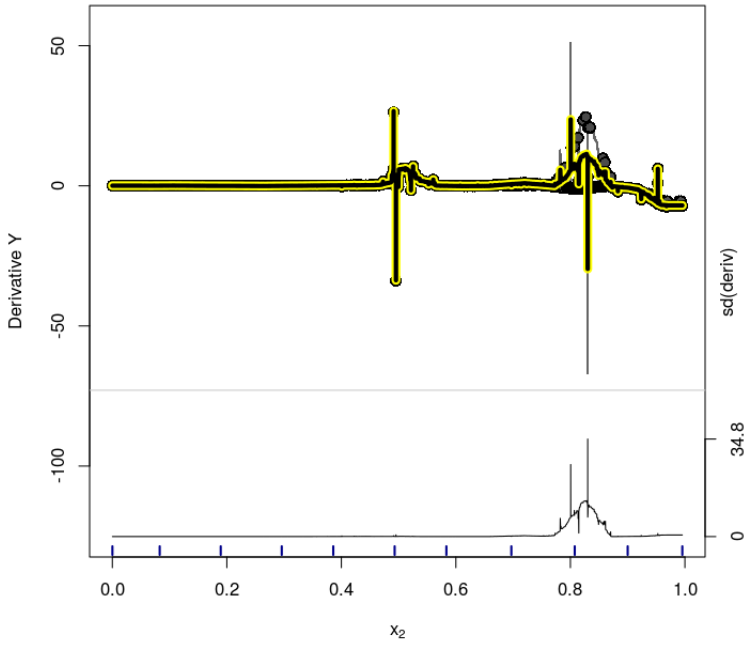

Figure D.146: *PS-DART model (Friedman Function under Sparsity) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

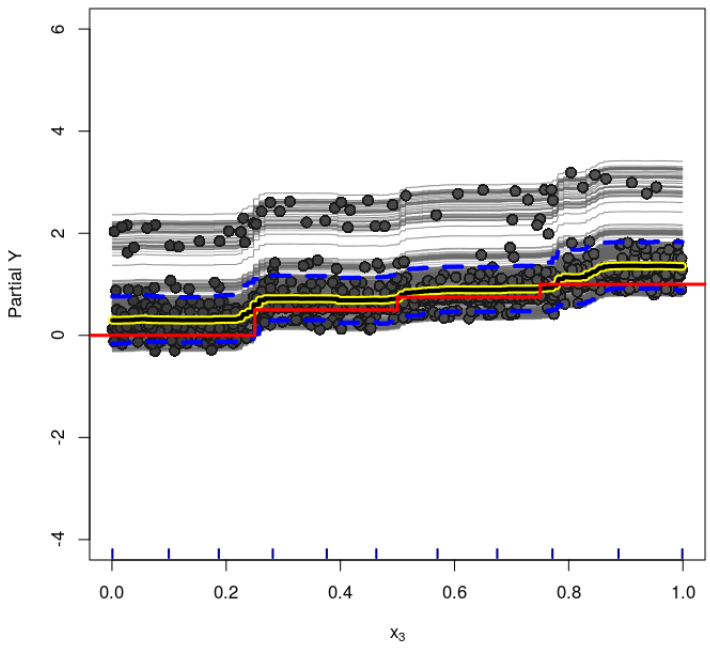

Figure D.147: *PS-DART model (Friedman Function under Sparsity) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

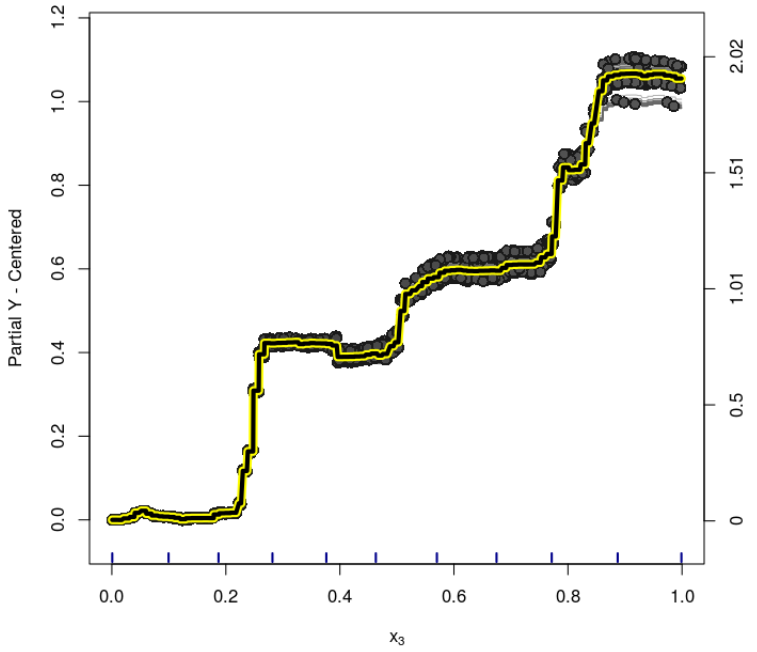

Figure D.148: *PS-DART model (Friedman Function under Sparsity) - Variable x_3 - centered-ICE Plot for the treatment effect.*

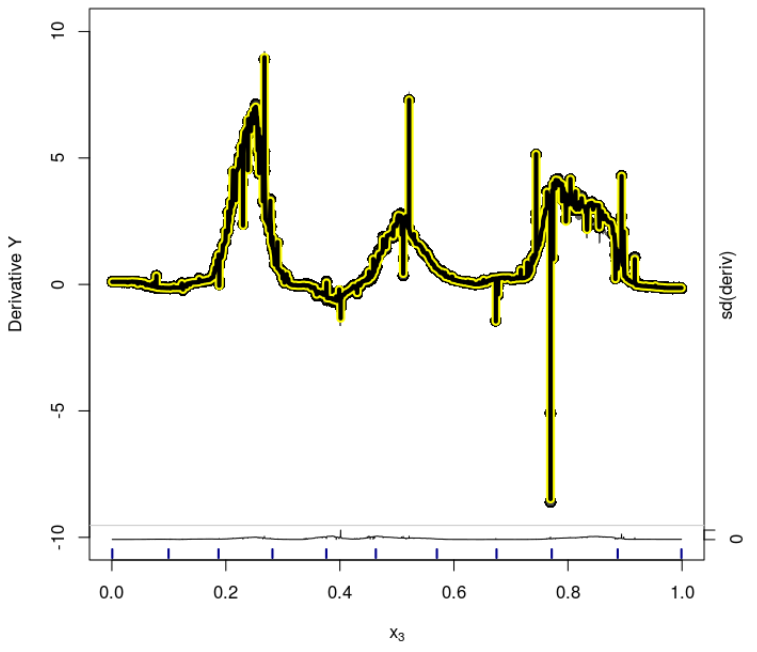

Figure D.149: *PS-DART model (Friedman Function under Sparsity) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

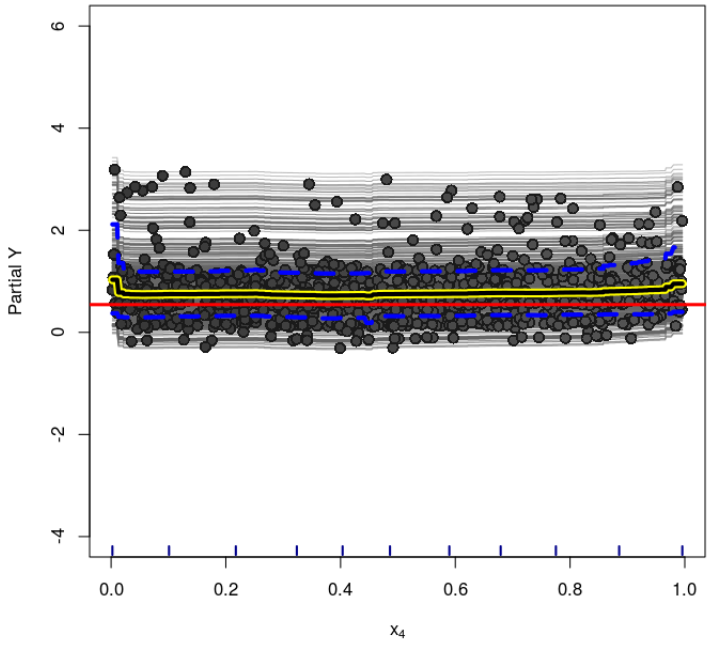

Figure D.150: *PS-DART model (Friedman Function under Sparsity) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

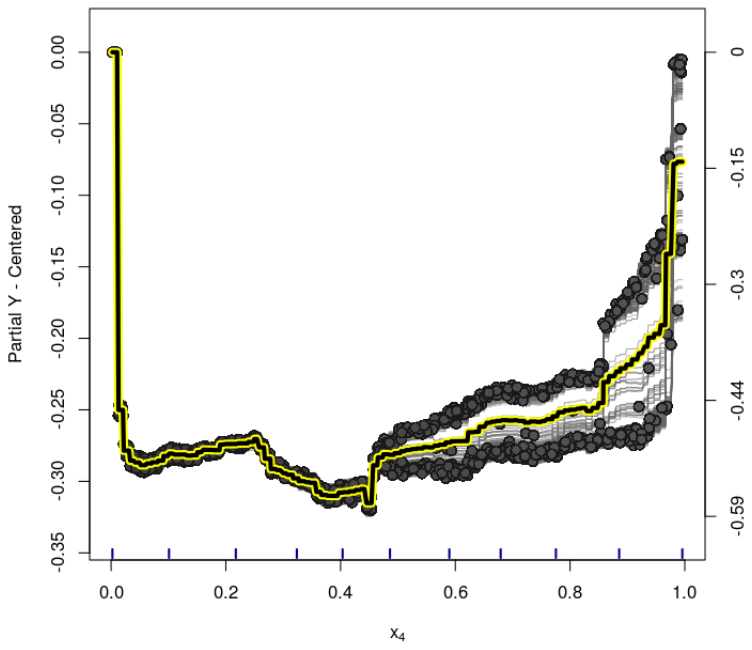

Figure D.151: *PS-DART model (Friedman Function under Sparsity) - Variable x_4 - centered-ICE Plot for the treatment effect.*

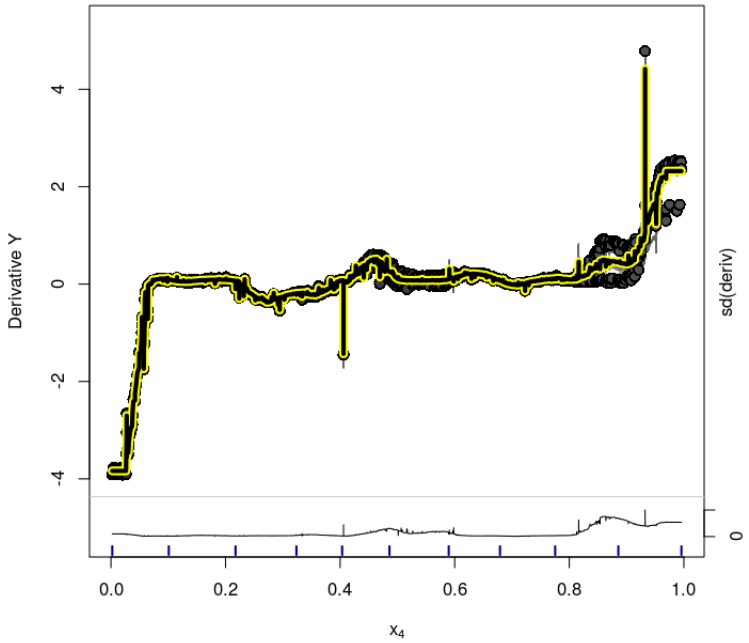

Figure D.152: *PS-DART model (Friedman Function under Sparsity) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

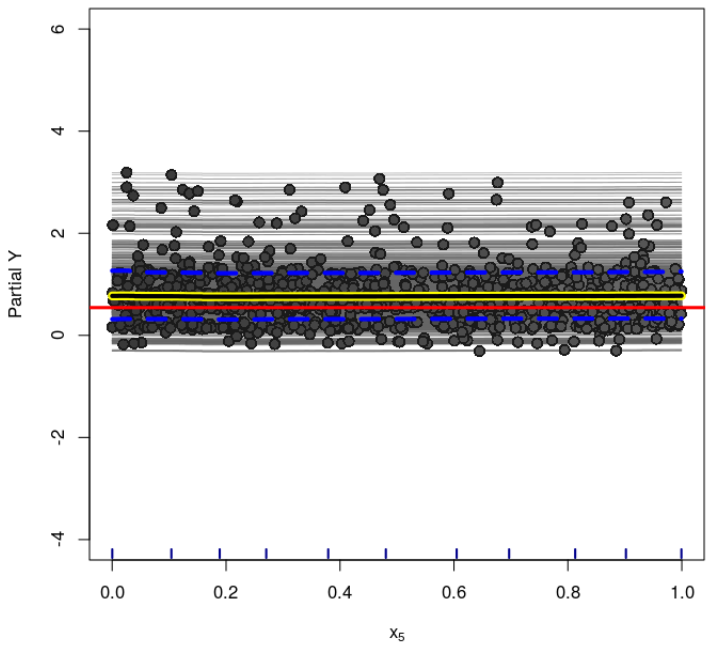

Figure D.153: *PS-DART model (Friedman Function under Sparsity) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

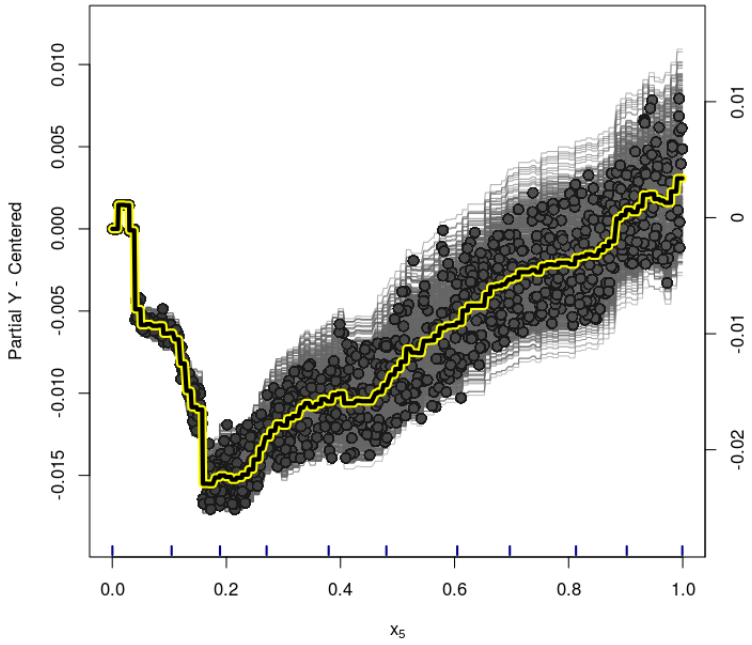

Figure D.154: *PS-DART model (Friedman Function under Sparsity) - Variable x_5 - centered-ICE Plot for the treatment effect.*

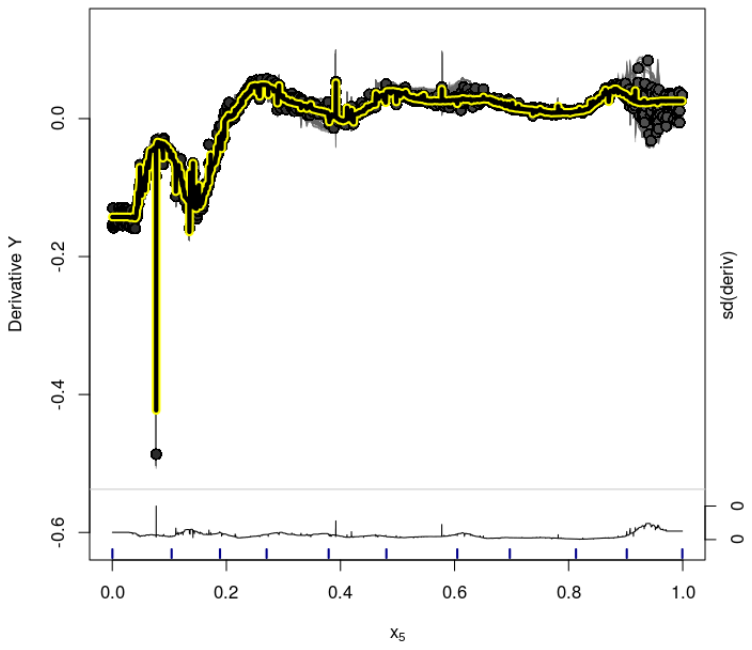

Figure D.155: *PS-DART model (Friedman Function under Sparsity) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

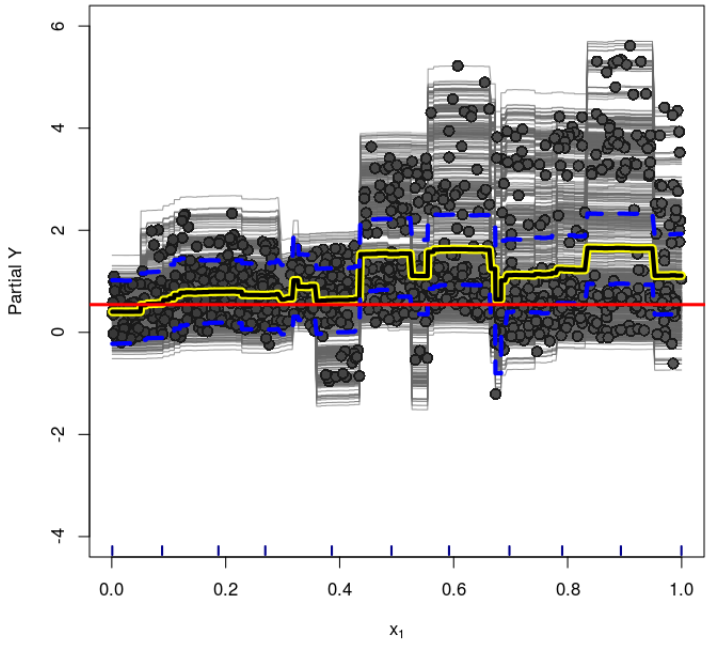

Figure D.156: *GLM-DART model (Friedman Function under Sparsity) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

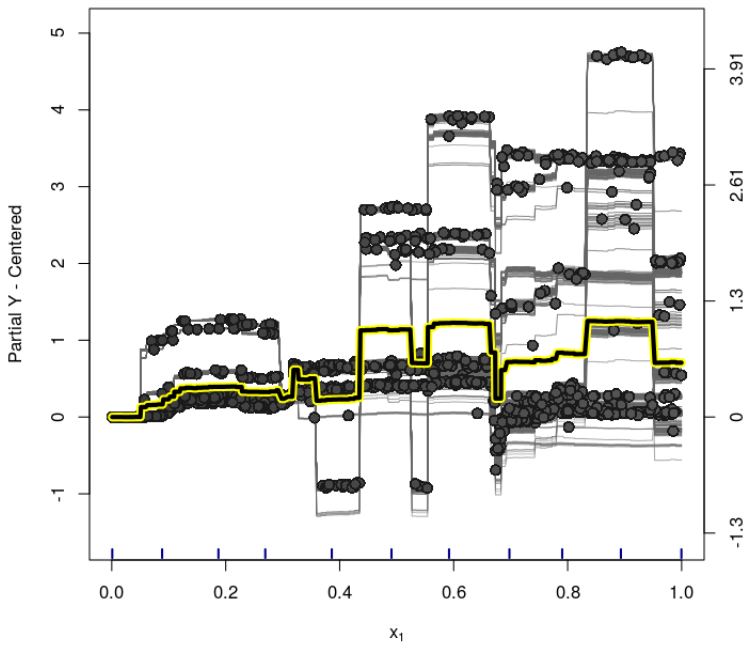

Figure D.157: *GLM-DART model (Friedman Function under Sparsity) - Variable x_1 - centered-ICE Plot for the treatment effect.*

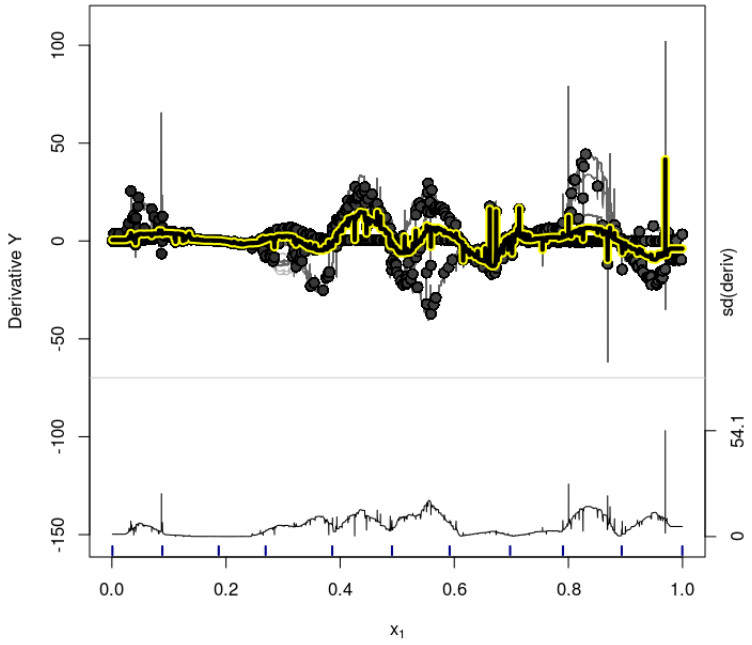

Figure D.158: *GLM-DART model (Friedman Function under Sparsity) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

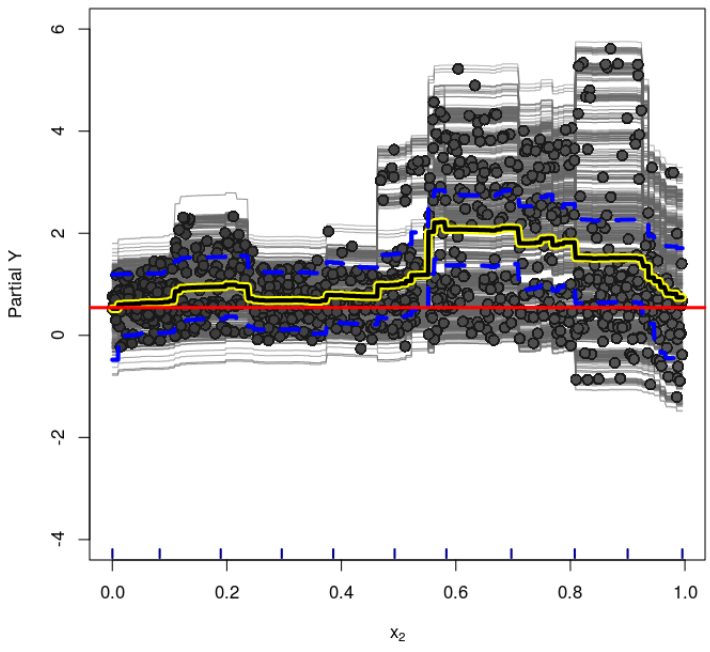

Figure D.159: *GLM-DART model (Friedman Function under Sparsity) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

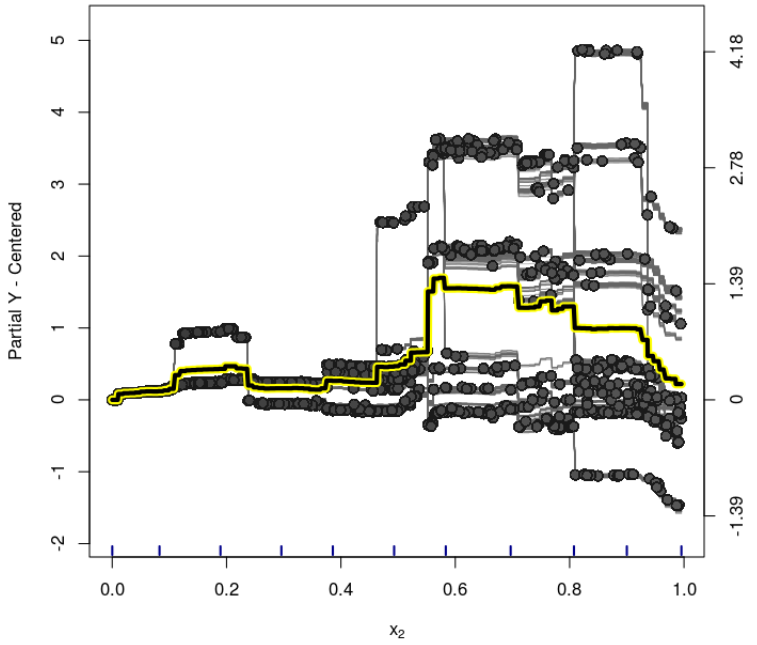

Figure D.160: *GLM-DART model (Friedman Function under Sparsity) - Variable x_2 - centered-ICE Plot for the treatment effect.*

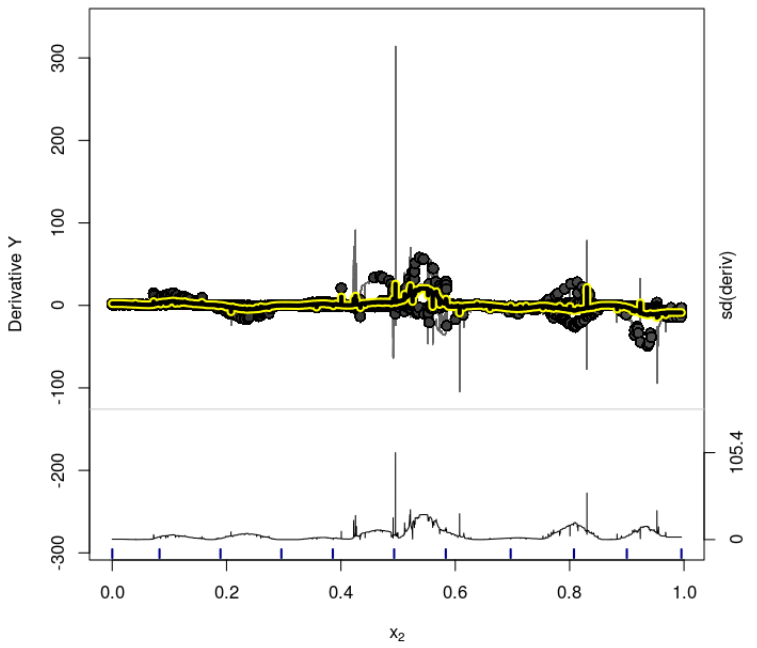

Figure D.161: *GLM-DART model (Friedman Function under Sparsity) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

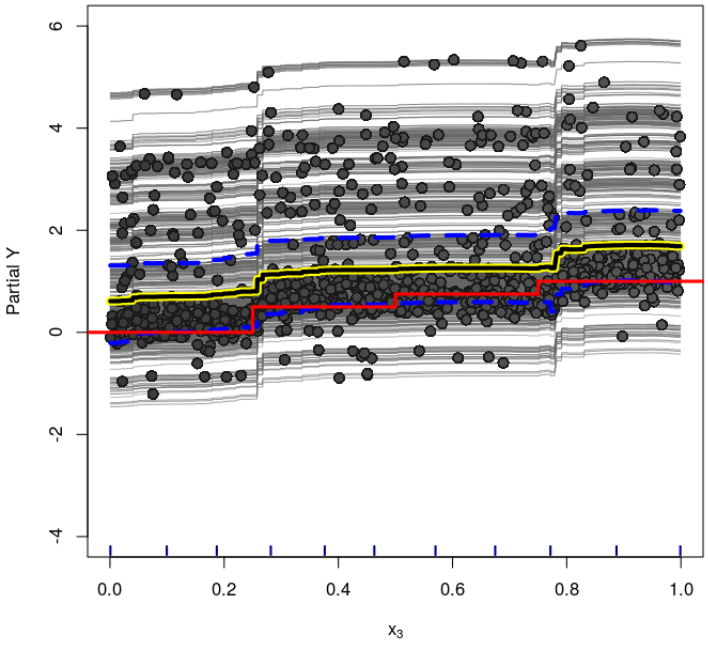

Figure D.162: *GLM-DART model (Friedman Function under Sparsity) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

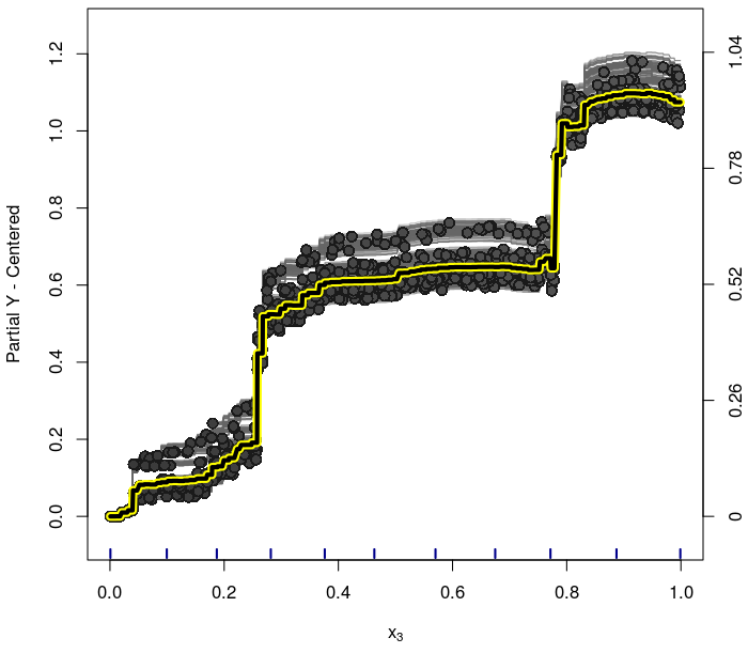

Figure D.163: *GLM-DART model (Friedman Function under Sparsity) - Variable x_3 - centered-ICE Plot for the treatment effect.*

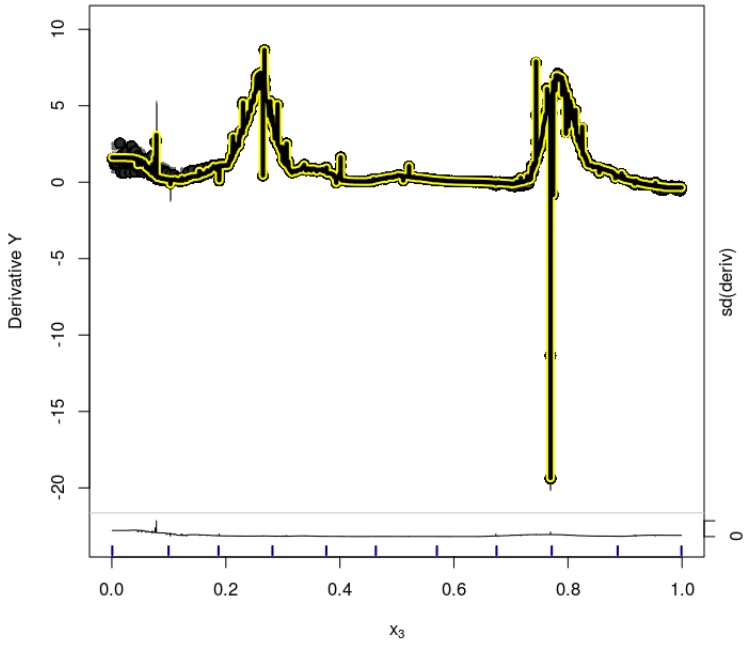

Figure D.164: *GLM-DART model (Friedman Function under Sparsity) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

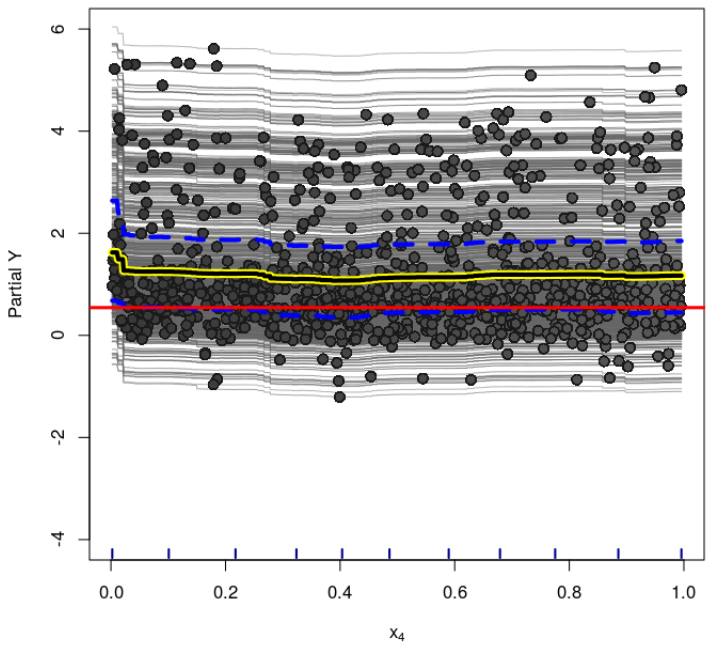

Figure D.165: *GLM-DART model (Friedman Function under Sparsity) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

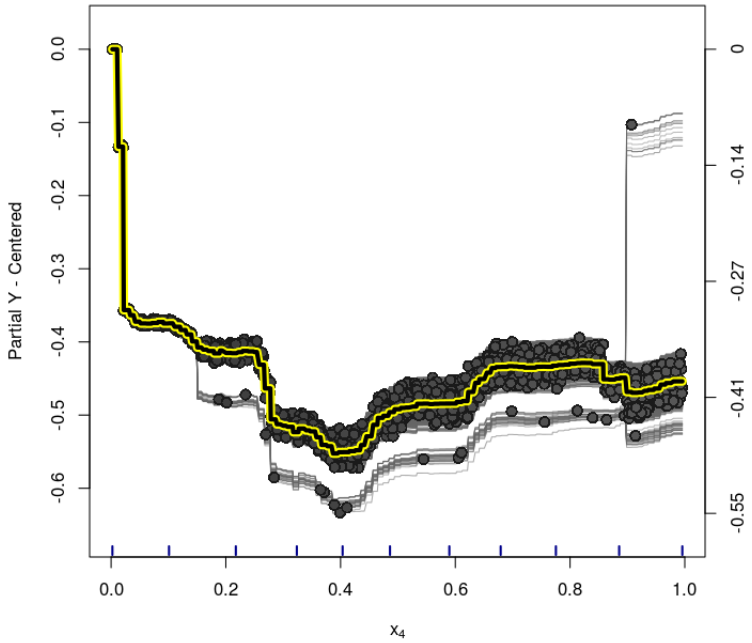

Figure D.166: *GLM-DART model (Friedman Function under Sparsity) - Variable x_4 - centered-ICE Plot for the treatment effect.*

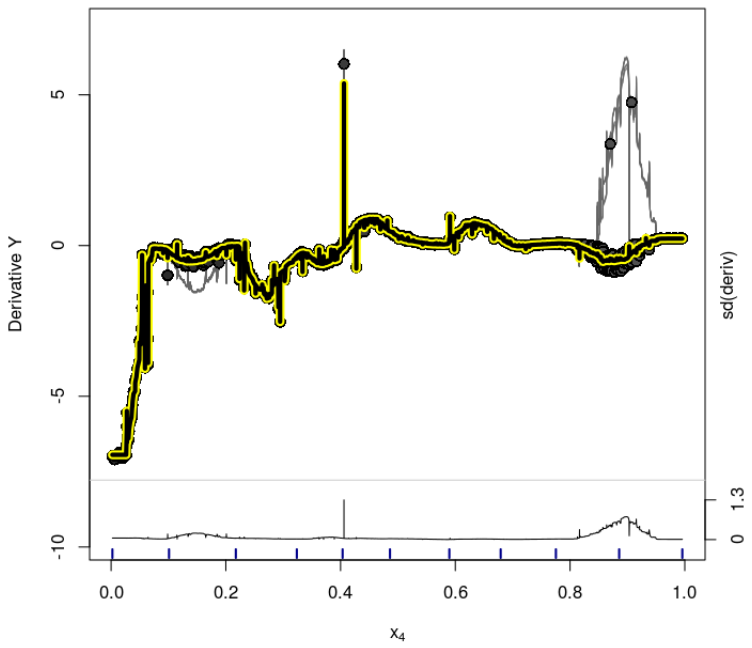

Figure D.167: *GLM-DART model (Friedman Function under Sparsity) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

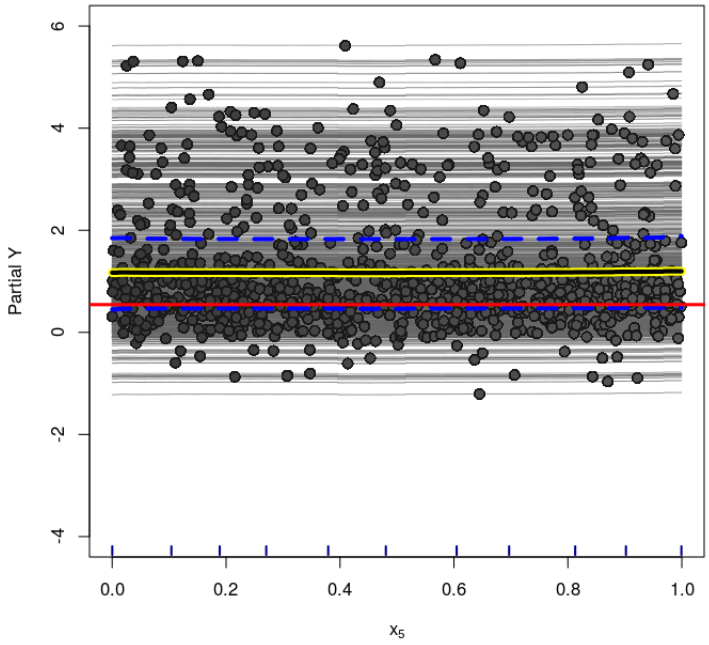

Figure D.168: *GLM-DART model (Friedman Function under Sparsity) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

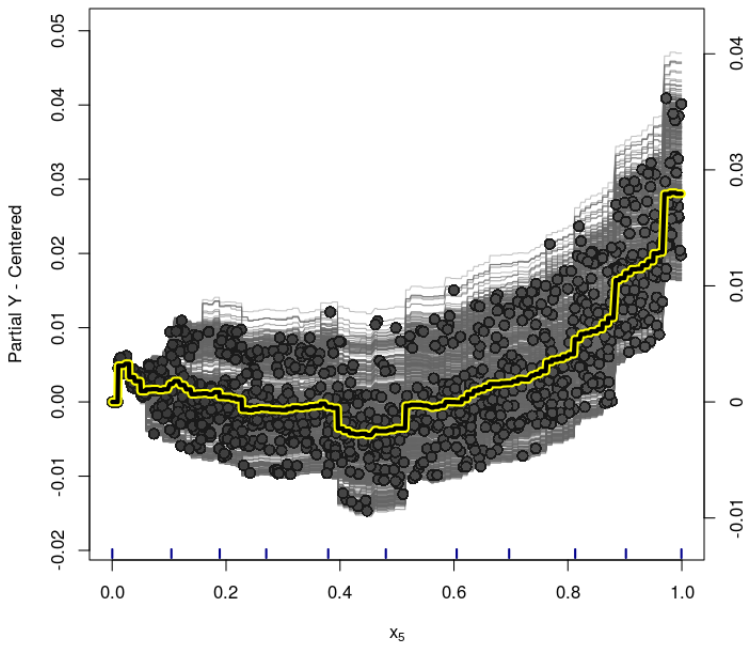

Figure D.169: *GLM-DART model (Friedman Function under Sparsity) - Variable x_5 - centered-ICE Plot for the treatment effect.*

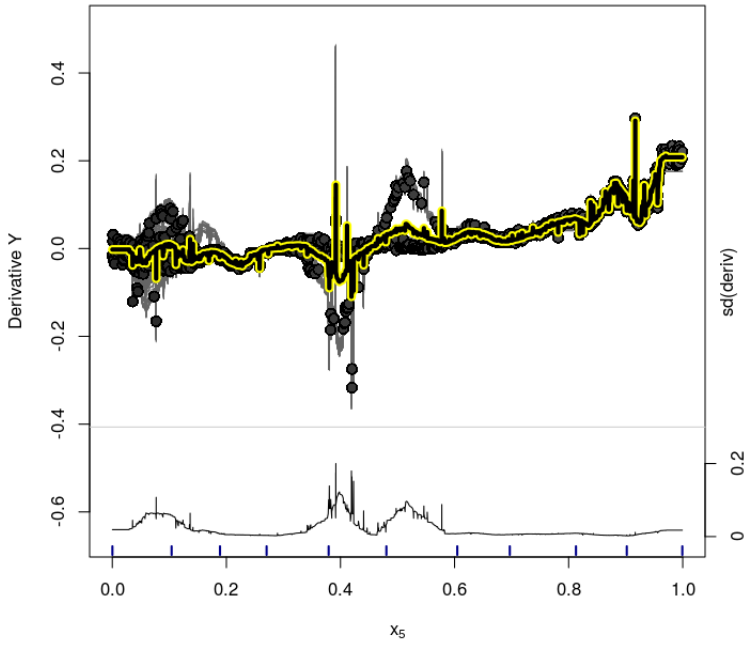

Figure D.170: *GLM-DART model (Friedman Function under Sparsity) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

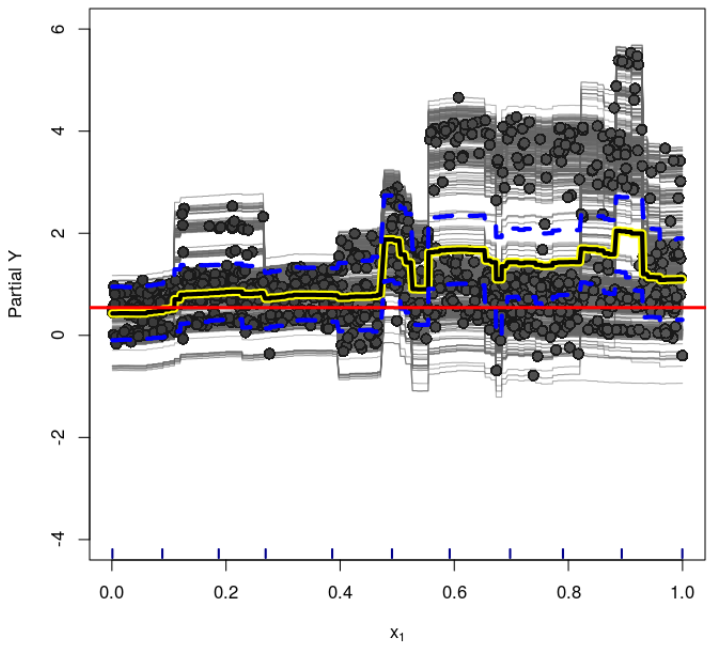

Figure D.171: *Rand-DART model (Friedman Function under Sparsity) - Variable x_1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

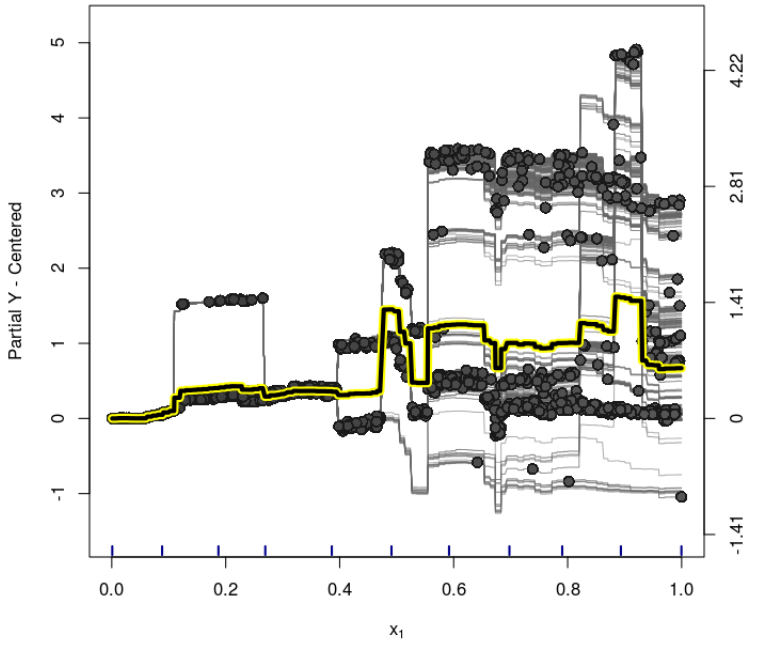

Figure D.172: *Rand-DART model (Friedman Function under Sparsity) - Variable x_1 - centered-ICE Plot for the treatment effect.*

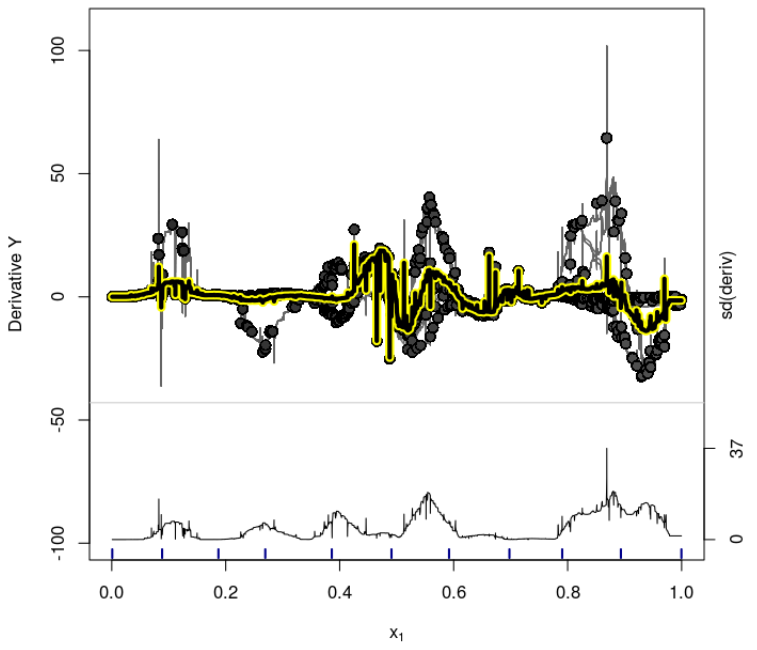

Figure D.173: *Rand-DART model (Friedman Function under Sparsity) - Variable x_1 - d-ICE Plot for the treatment effect estimates.*

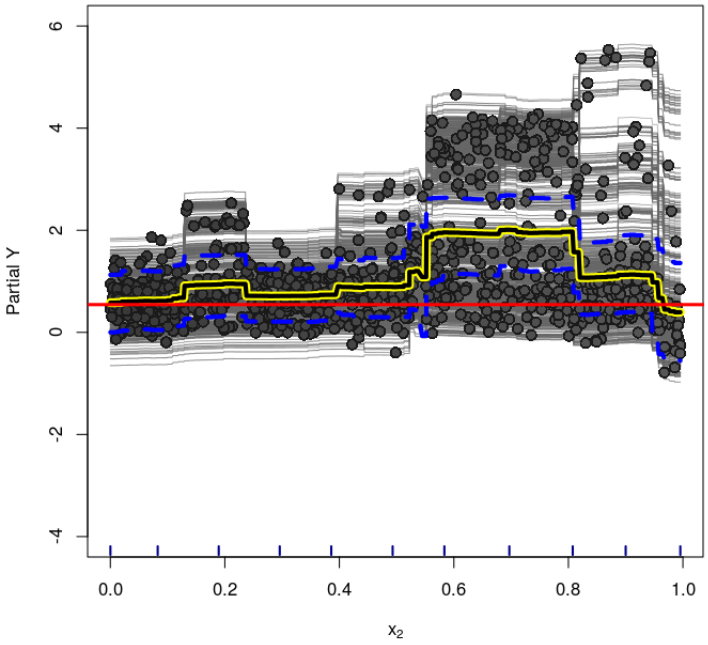

Figure D.174: *Rand-DART model (Friedman Function under Sparsity) - Variable x_2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

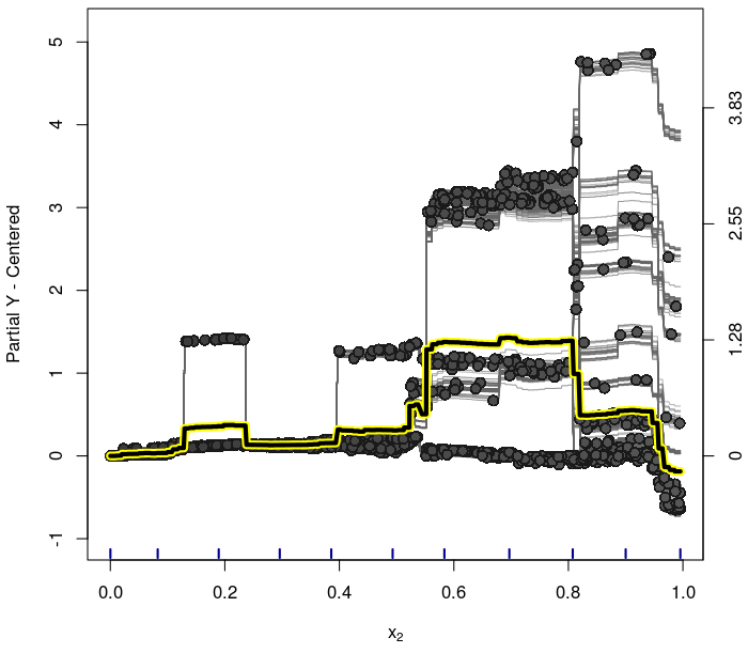

Figure D.175: *Rand-DART model (Friedman Function under Sparsity) - Variable x_2 - centered-ICE Plot for the treatment effect.*

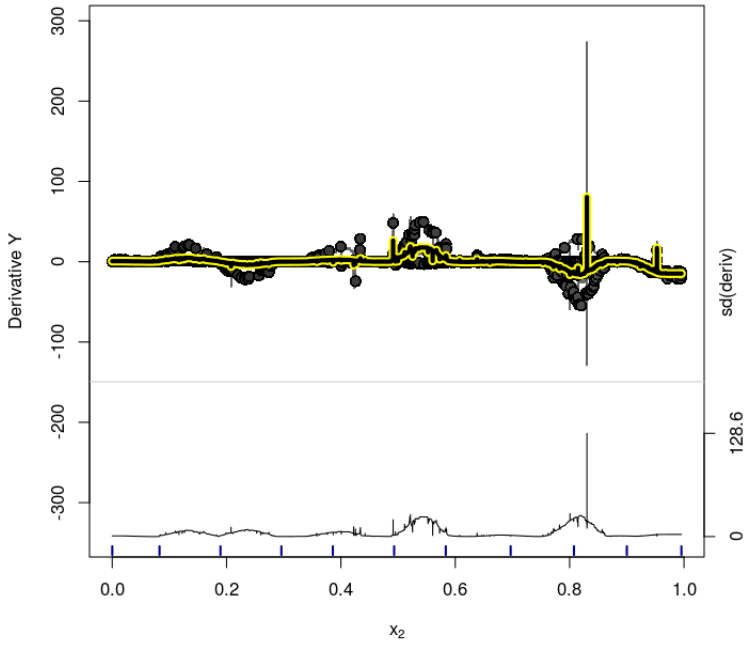

Figure D.176: *Rand-DART model (Friedman Function under Sparsity) - Variable x_2 - d-ICE Plot for the treatment effect estimates.*

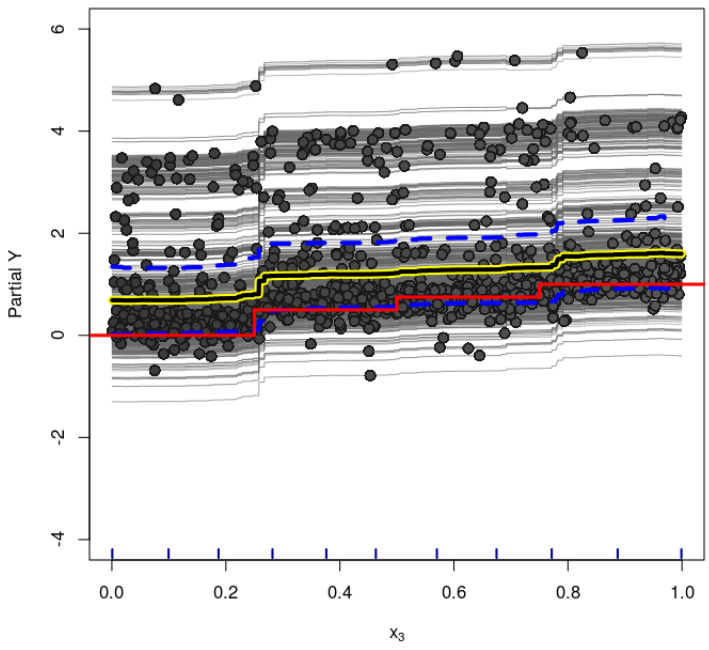

Figure D.177: *Rand-DART model (Friedman Function under Sparsity) - Variable x_3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

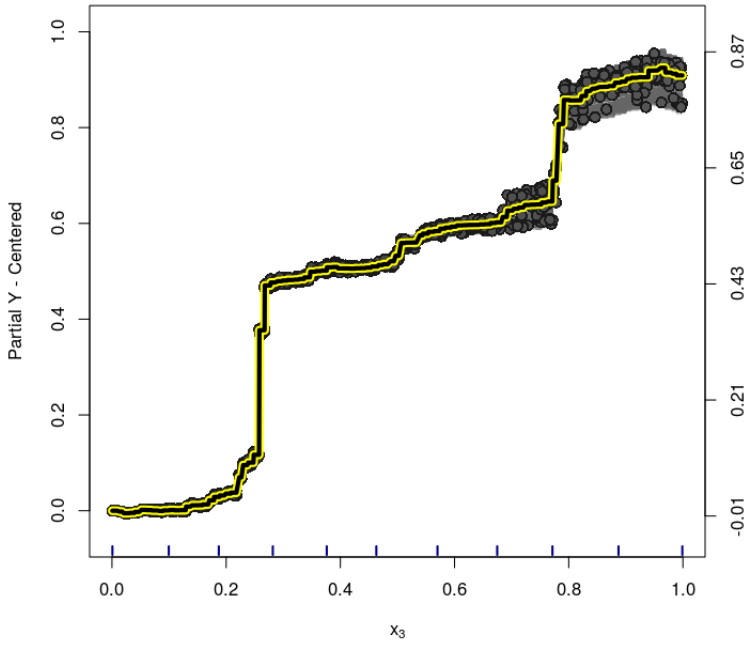

Figure D.178: *Rand-DART model (Friedman Function under Sparsity) - Variable x_3 - centered-ICE Plot for the treatment effect.*

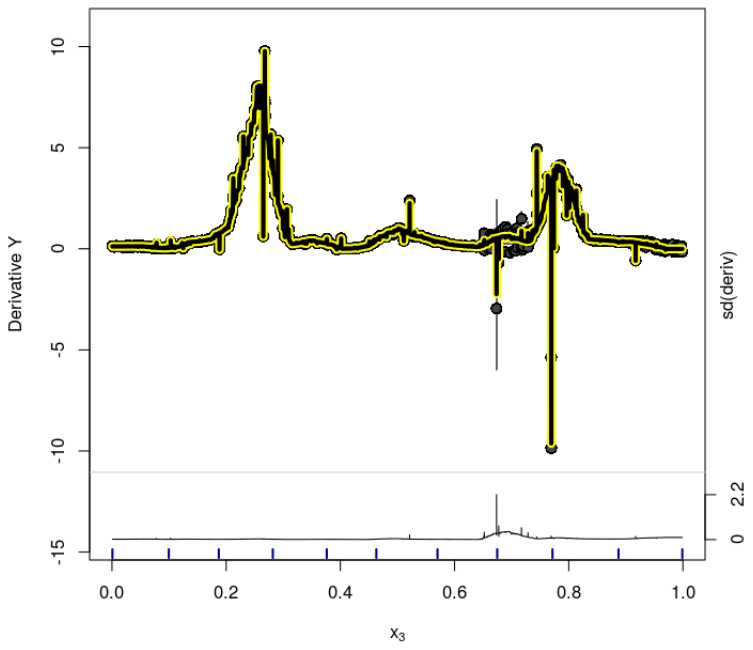

Figure D.179: *Rand-DART model (Friedman Function under Sparsity) - Variable x_3 - d-ICE Plot for the treatment effect estimates.*

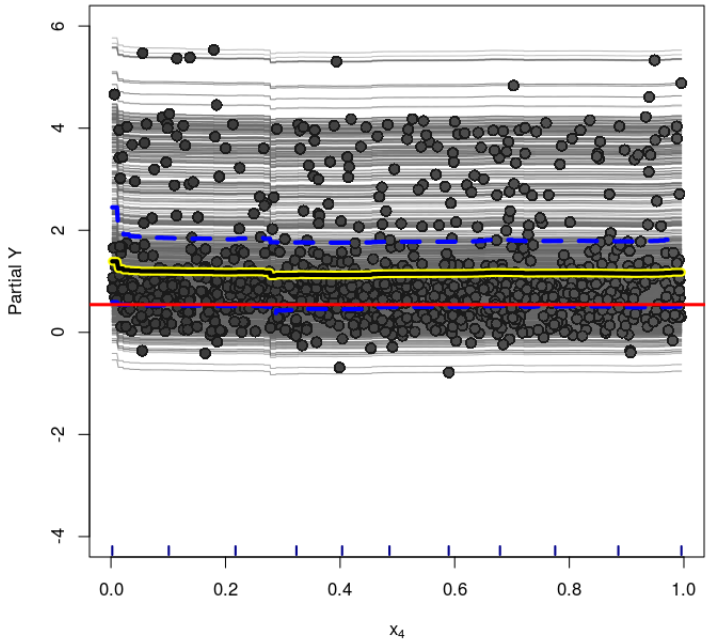

Figure D.180: *Rand-DART model (Friedman Function under Sparsity) - Variable x_4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

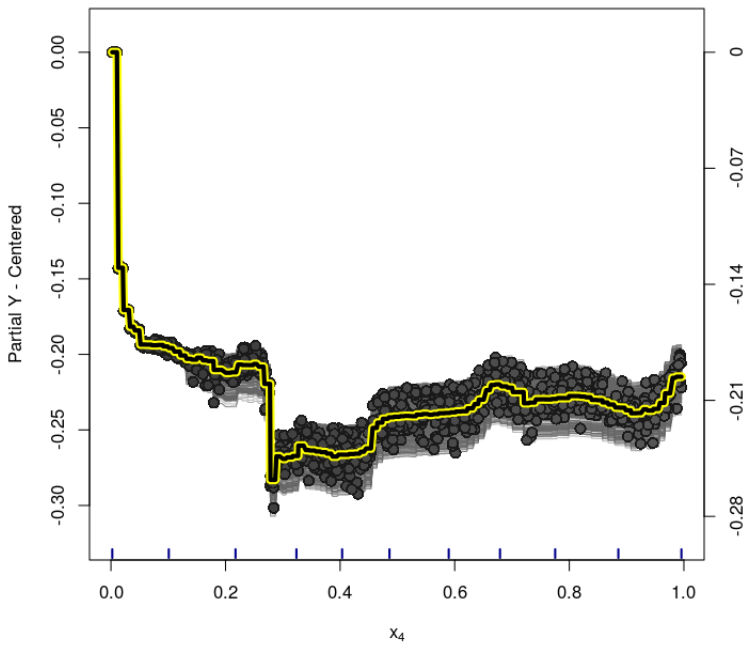

Figure D.181: *Rand-DART model (Friedman Function under Sparsity) - Variable x_4 - centered-ICE Plot for the treatment effect.*

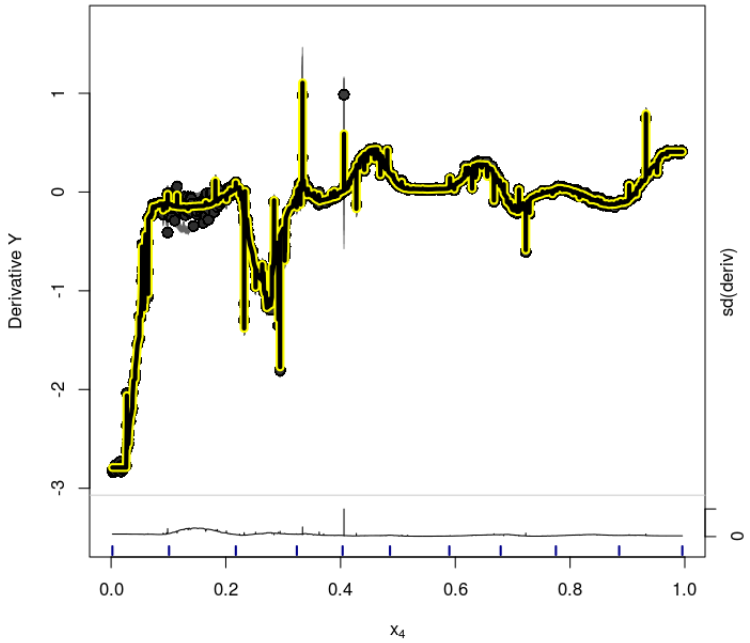

Figure D.182: *Rand-DART model (Friedman Function under Sparsity) - Variable x_4 - d-ICE Plot for the treatment effect estimates.*

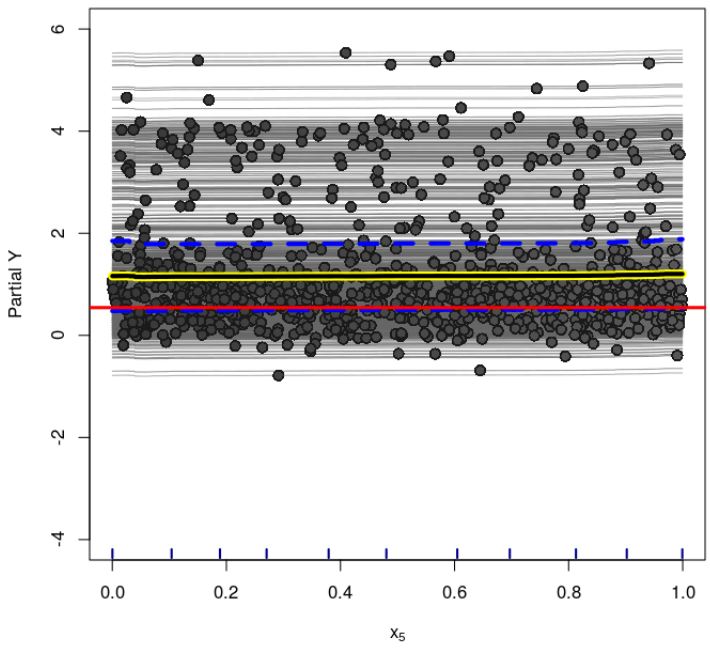

Figure D.183: *Rand-DART model (Friedman Function under Sparsity) - Variable x_5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

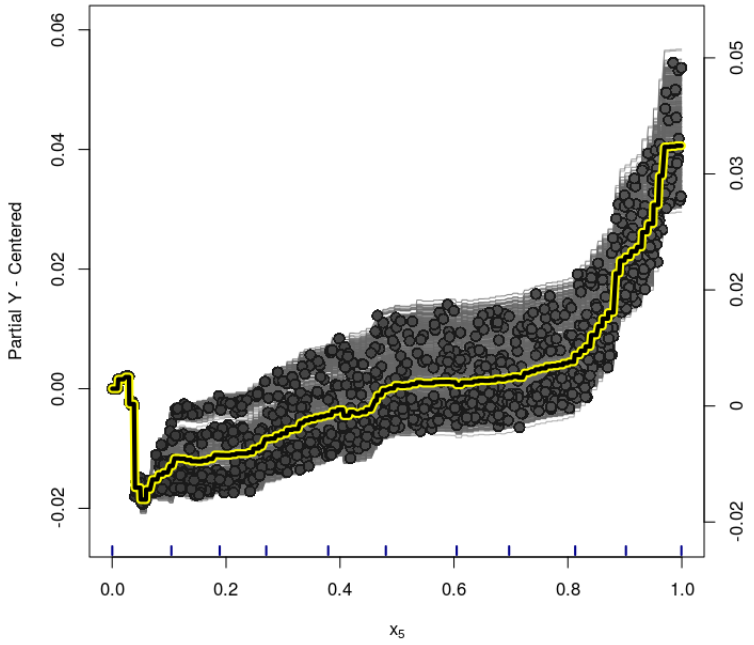

Figure D.184: *Rand-DART model (Friedman Function under Sparsity) - Variable x_5 - centered-ICE Plot for the treatment effect.*

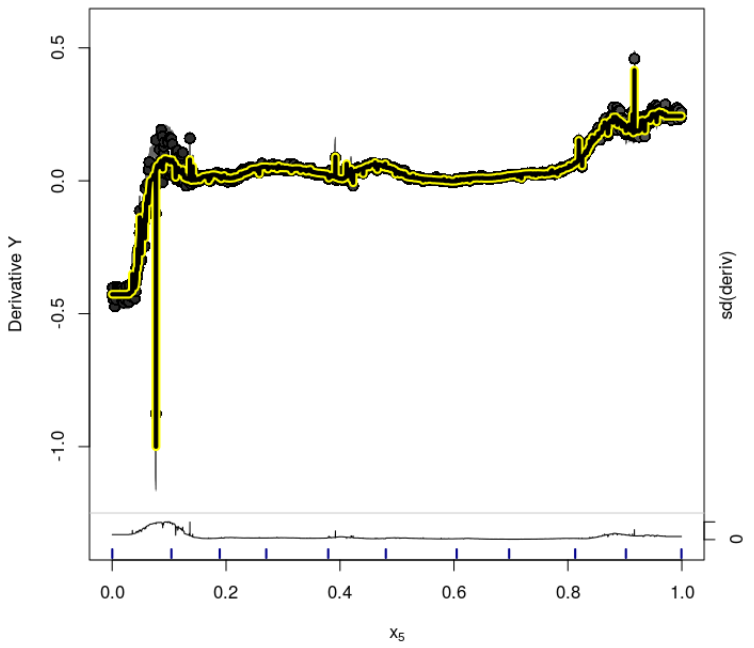

Figure D.185: *Rand-DART model (Friedman Function under Sparsity) - Variable x_5 - d-ICE Plot for the treatment effect estimates.*

Appendix E

Graphics - Real Data Analysis

E.1 DART Variable Selection

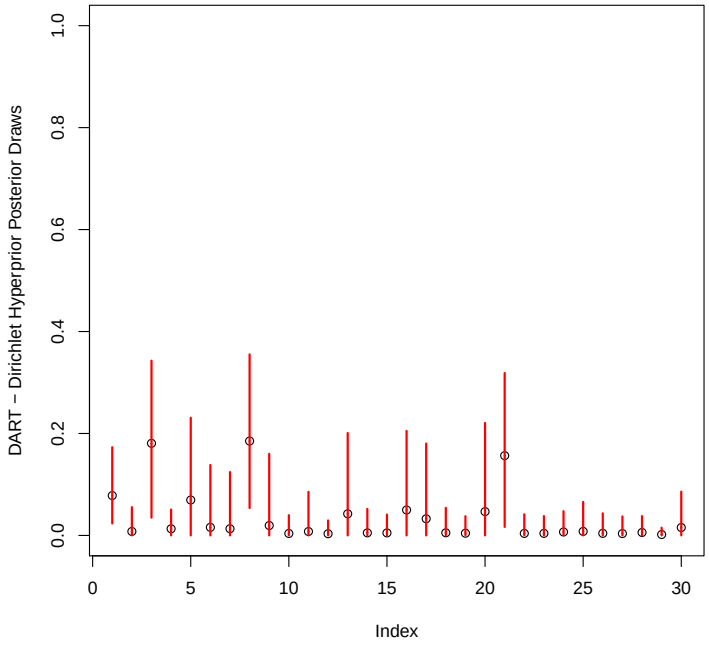

Figure E.1: *Real Data Analysis - Posterior draws from DART Dirichlet hyperprior with 95% credible intervals.*

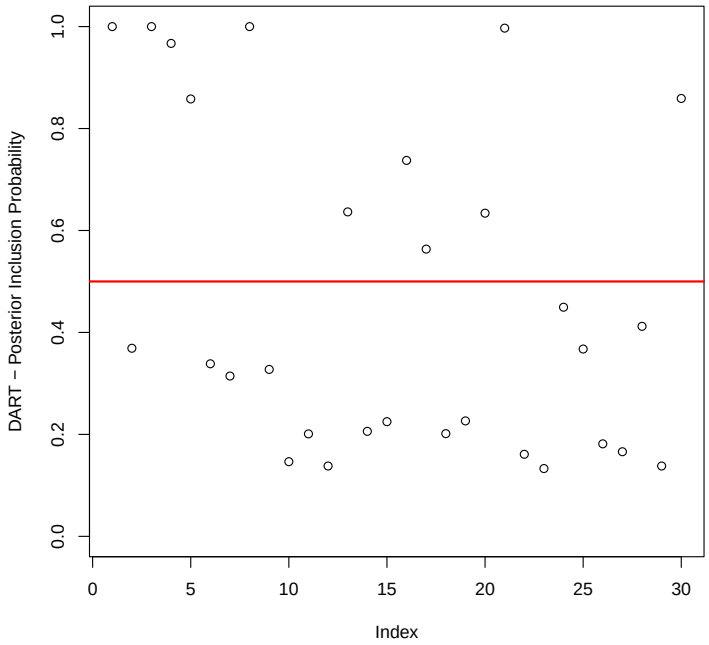

Figure E.2: *Real Data Analysis - Posterior Inclusion Probability of DART model.*

E.2 BCF ICE Plots

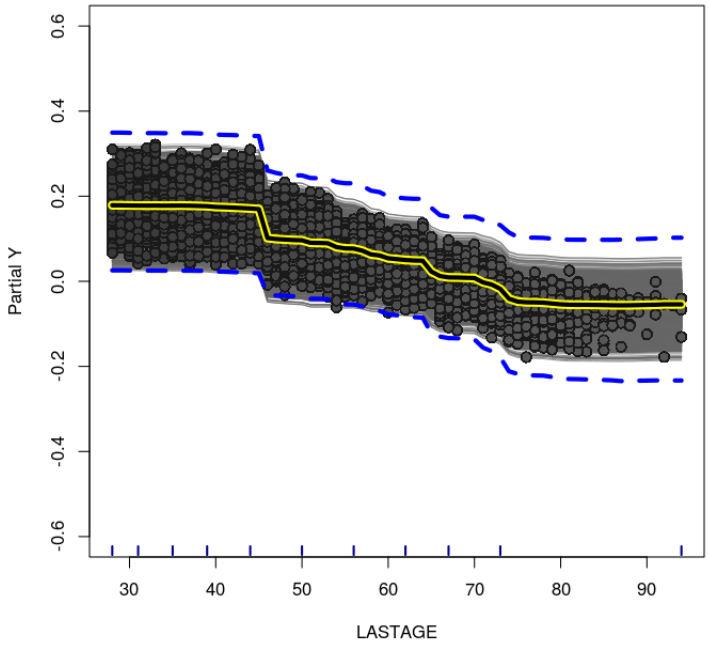

Figure E.3: *BCF model (Real Data Analysis) - Variable LASTAGE - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

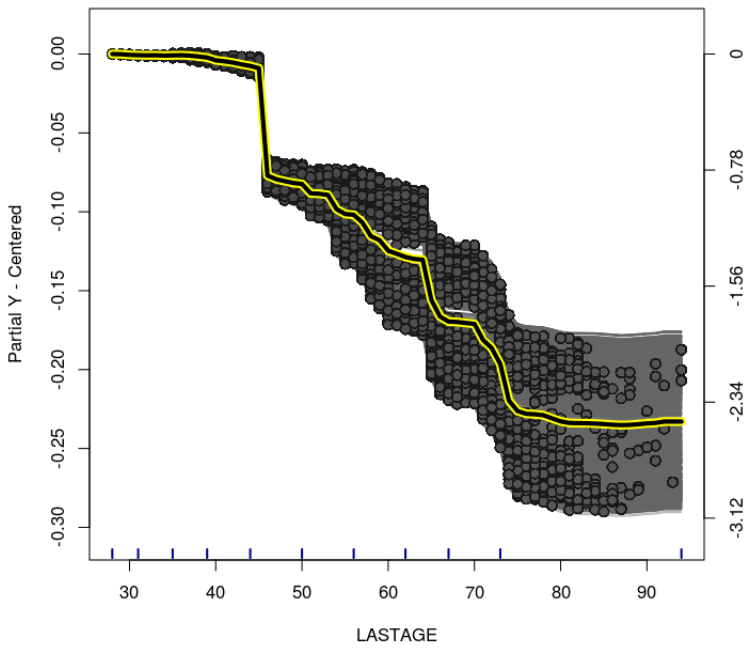

Figure E.4: *BCF model (Real Data Analysis) - Variable LASTAGE - centered-ICE Plot for the treatment effect.*

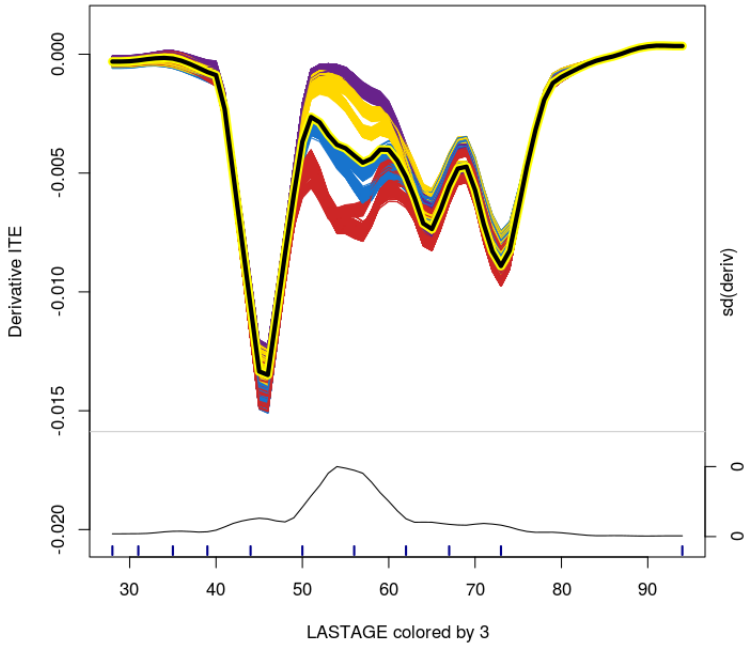

Figure E.5: *BCF model (Real Data Analysis) - Variable LASTAGE - d-ICE Plot for the treatment effect estimates.*

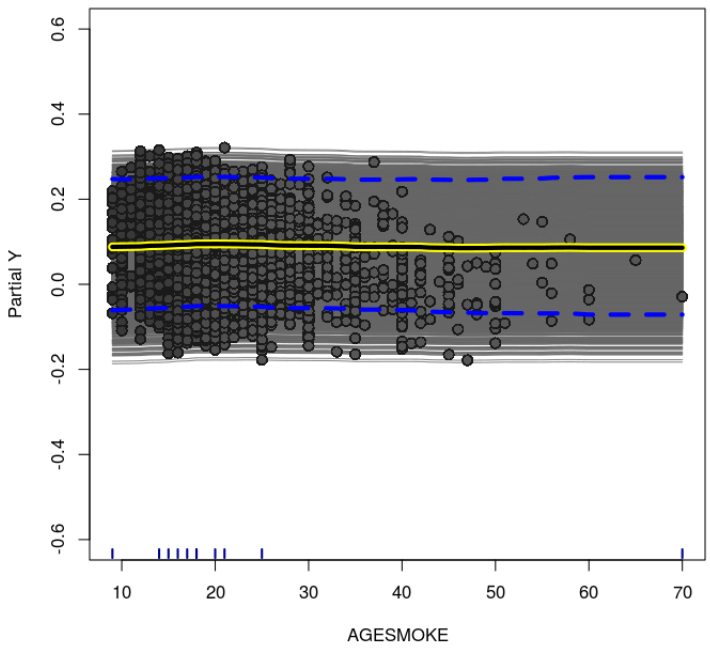

Figure E.6: *BCF model (Real Data Analysis) - Variable AGESMOKE - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

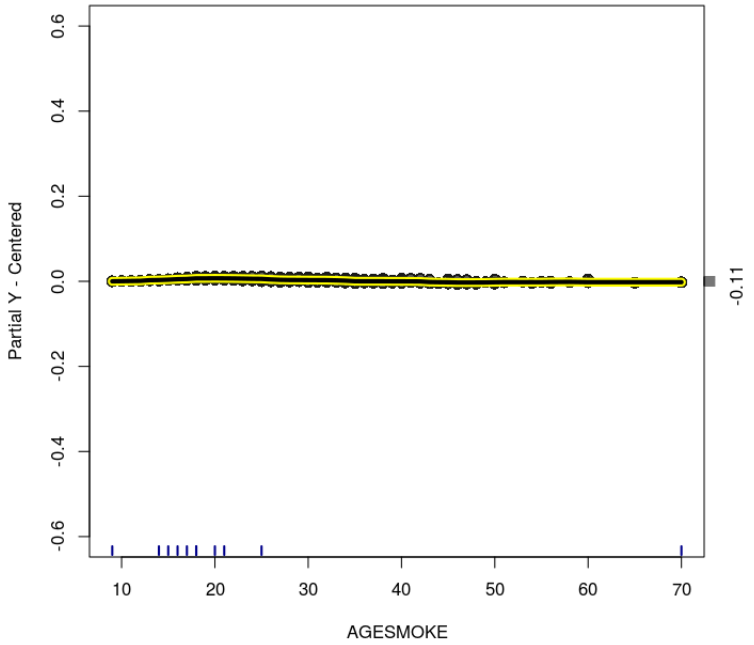

Figure E.7: *BCF model (Real Data Analysis) - Variable AGESMOKE - centered-ICE Plot for the treatment effect.*

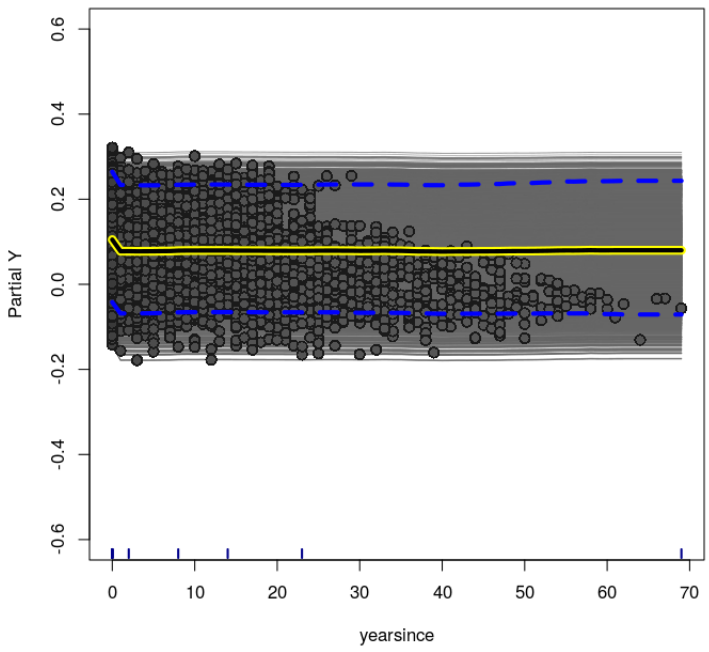

Figure E.8: *BCF model (Real Data Analysis) - Variable yearsince - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

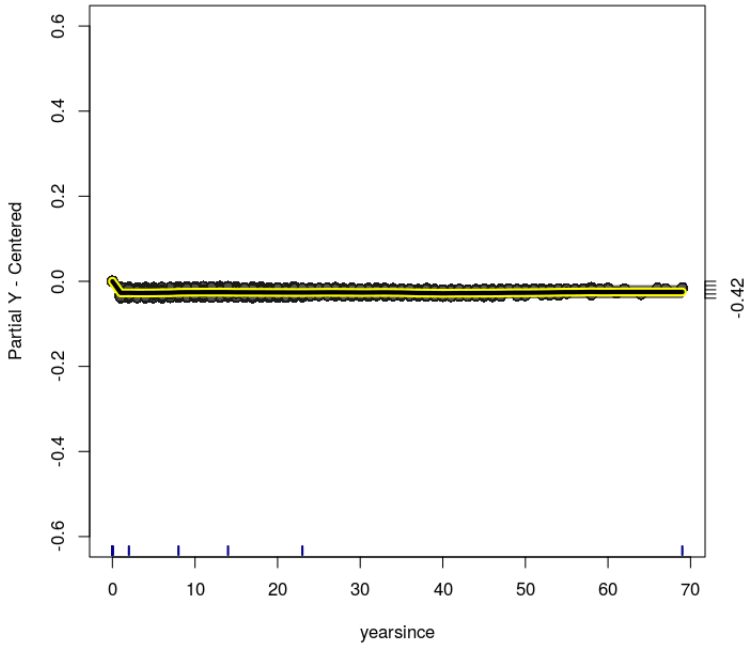

Figure E.9: *BCF model (Real Data Analysis) - Variable yearsince - centered-ICE Plot for the treatment effect.*

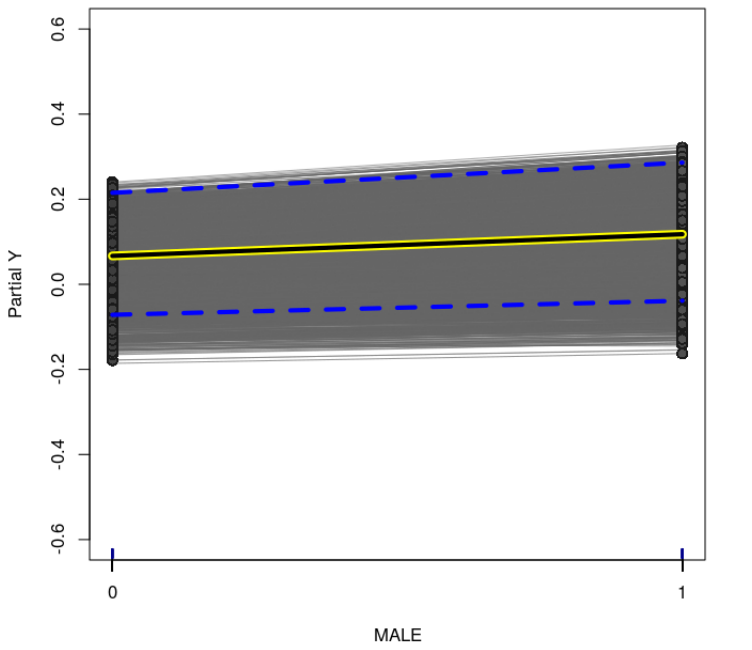

Figure E.10: *BCF model (Real Data Analysis) - Variable MALE - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

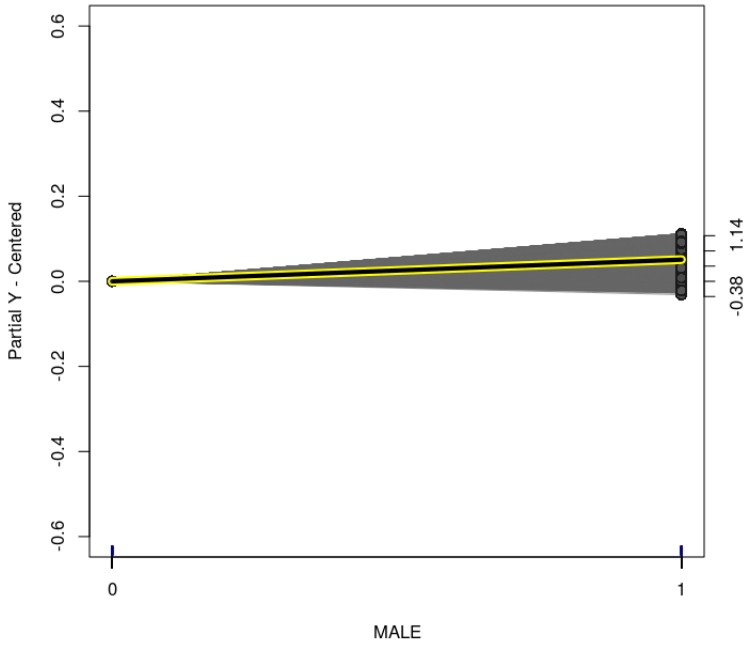

Figure E.11: *BCF model (Real Data Analysis) - Variable MALE - centered-ICE Plot for the treatment effect.*

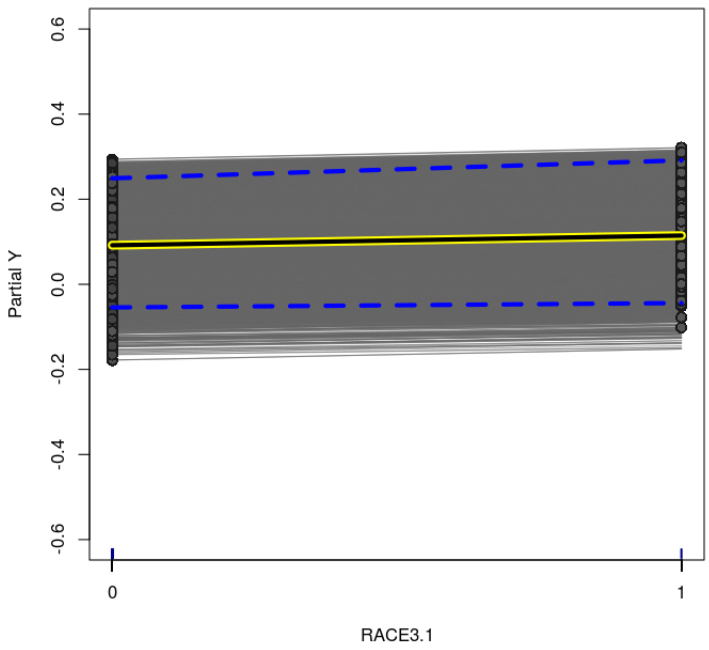

Figure E.12: *BCF model (Real Data Analysis) - Variable RACE3.1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

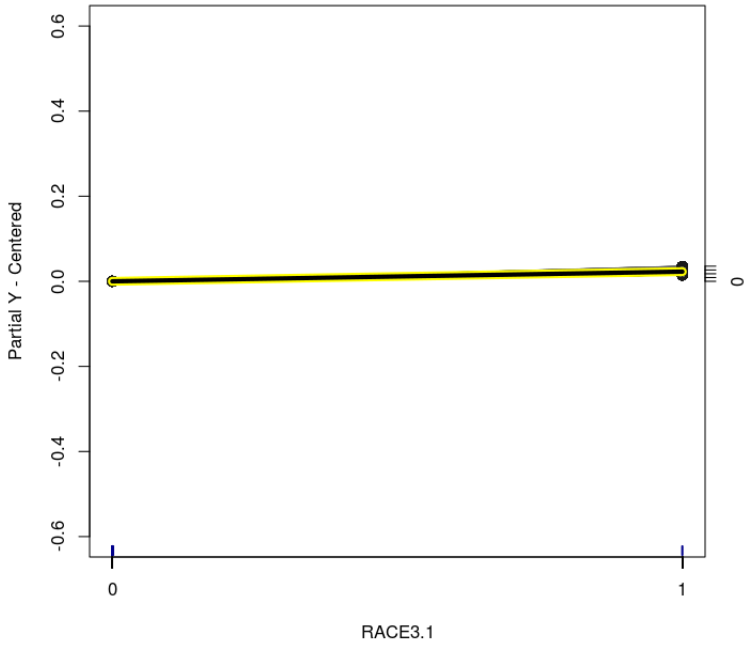

Figure E.13: *BCF model (Real Data Analysis) - Variable RACE*3.1 *- centered-ICE Plot for the treatment effect.*

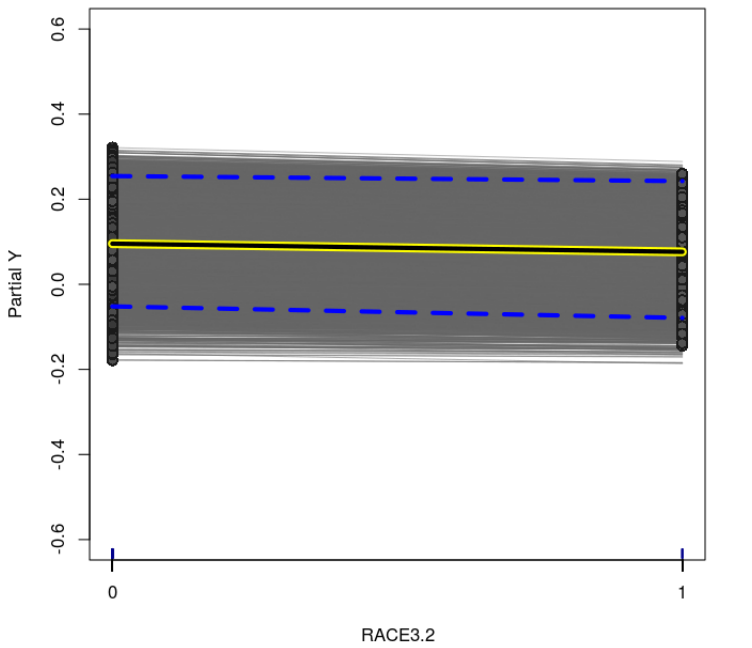

Figure E.14: *BCF model (Real Data Analysis) - Variable RACE*3.2 *- ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

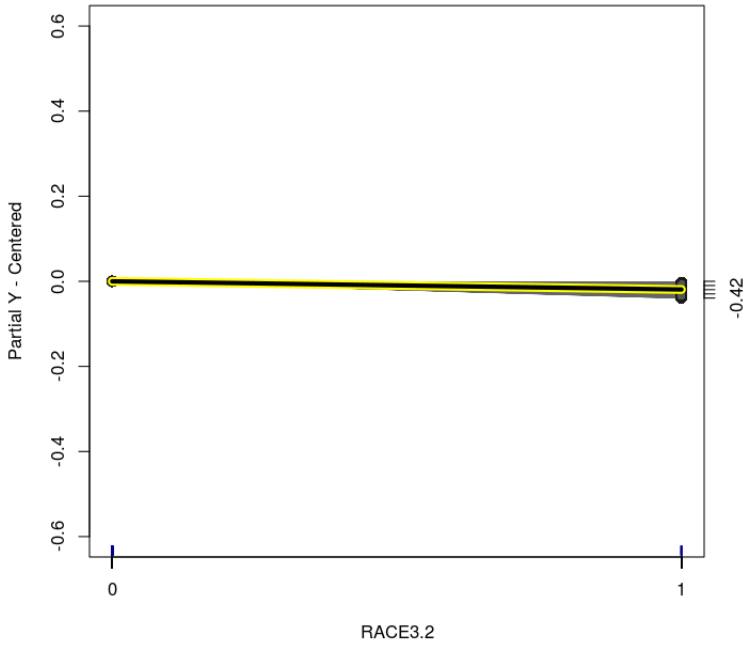

Figure E.15: *BCF model (Real Data Analysis) - Variable RACE*3.2 *- centered-ICE Plot for the treatment effect.*

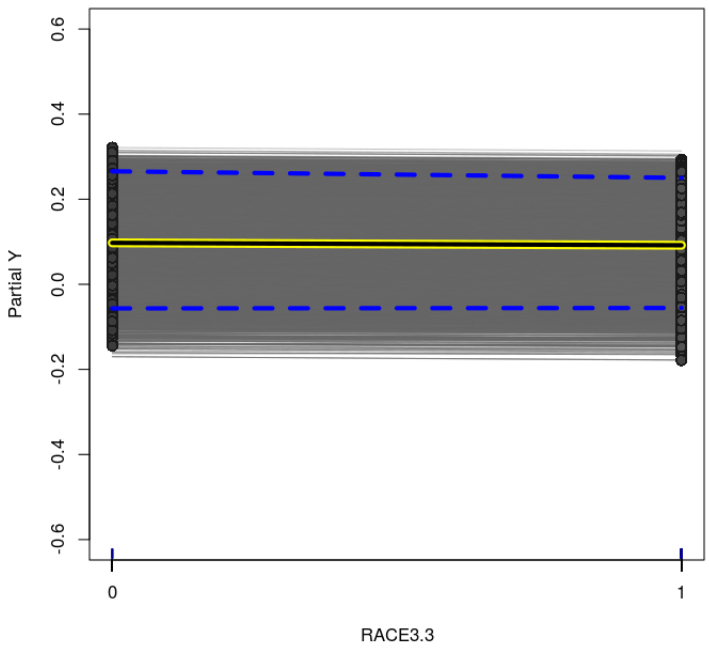

Figure E.16: *BCF model (Real Data Analysis) - Variable RACE*3.3 *- ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

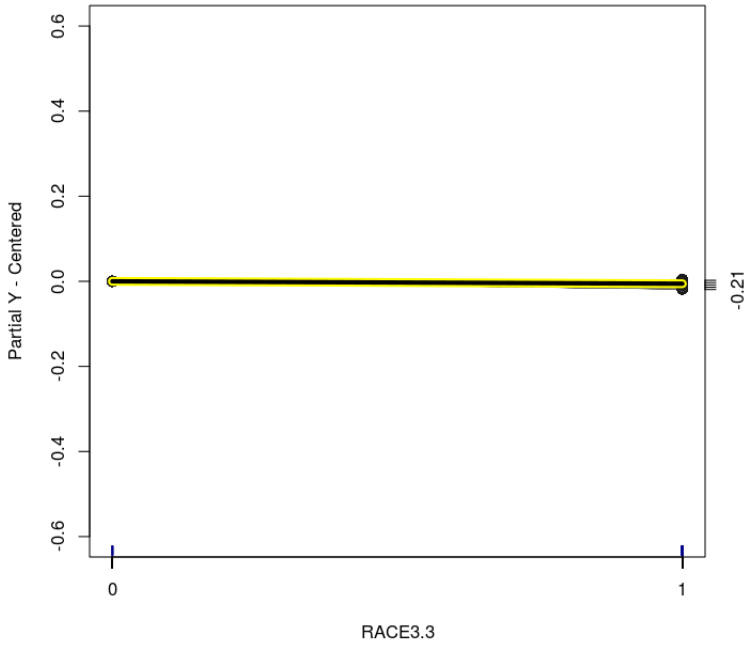

Figure E.17: *BCF model (Real Data Analysis) - Variable RACE*3.3 *- centered-ICE Plot for the treatment effect.*

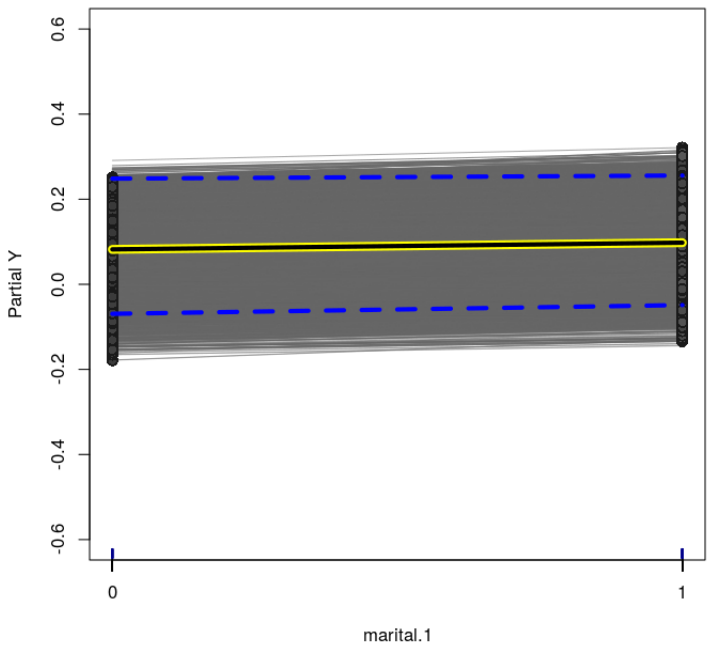

Figure E.18: *BCF model (Real Data Analysis) - Variable marital*.1 *- ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

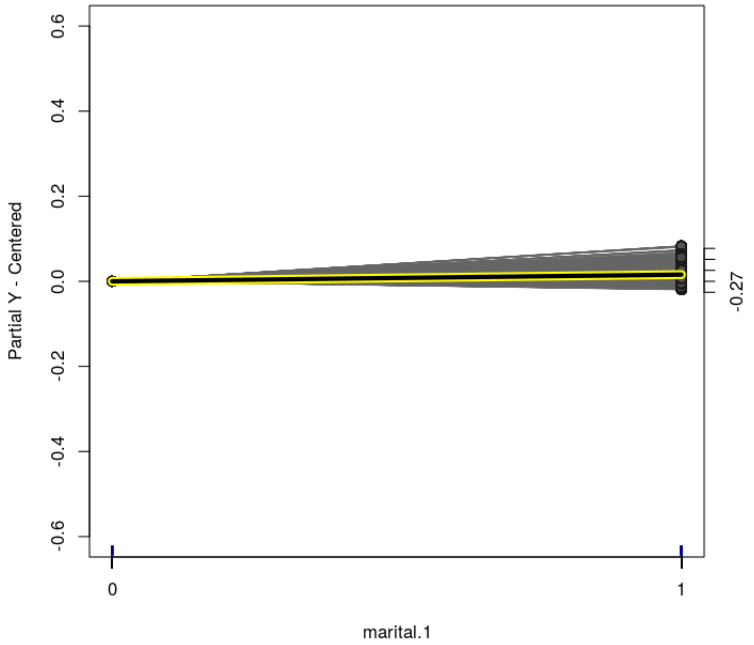

Figure E.19: *BCF model (Real Data Analysis) - Variable marital.*1 *- centered-ICE Plot for the treatment effect.*

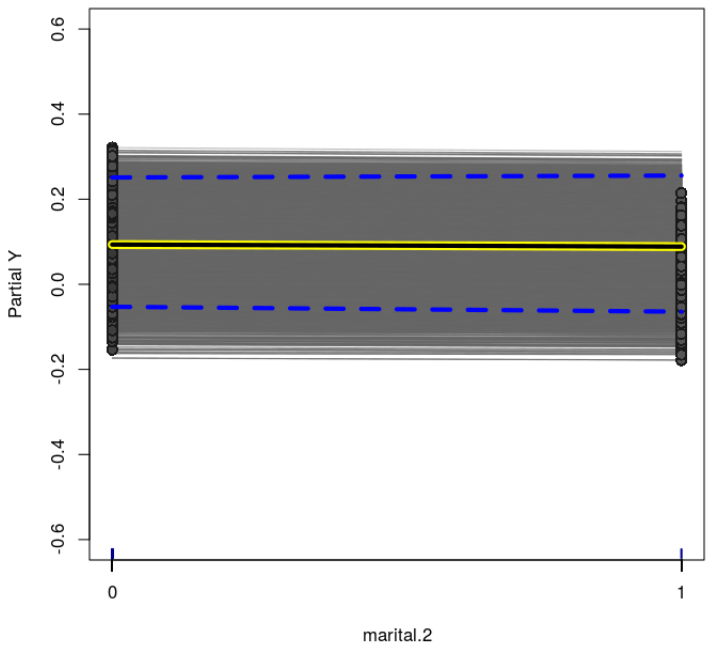

Figure E.20: *BCF model (Real Data Analysis) - Variable marital.*2 *- ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

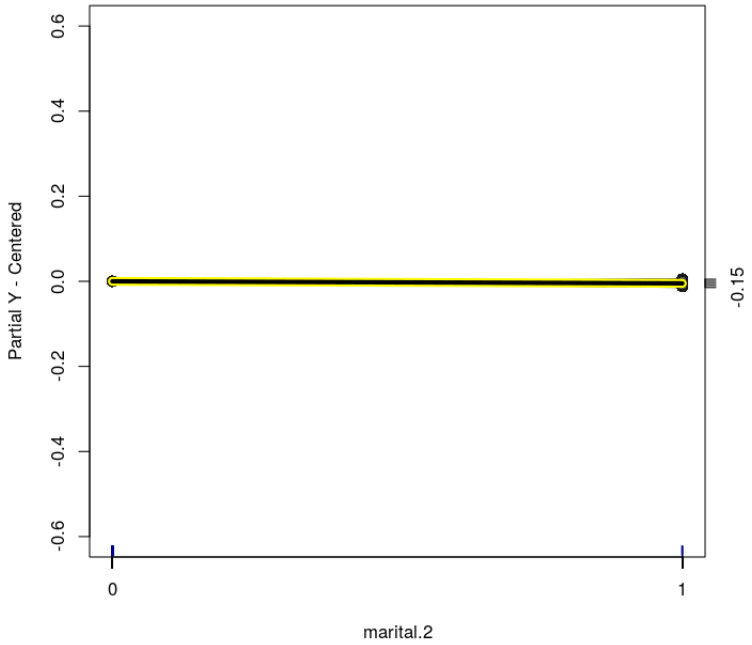

Figure E.21: *BCF model (Real Data Analysis) - Variable marital.2 - centered-ICE Plot for the treatment effect.*

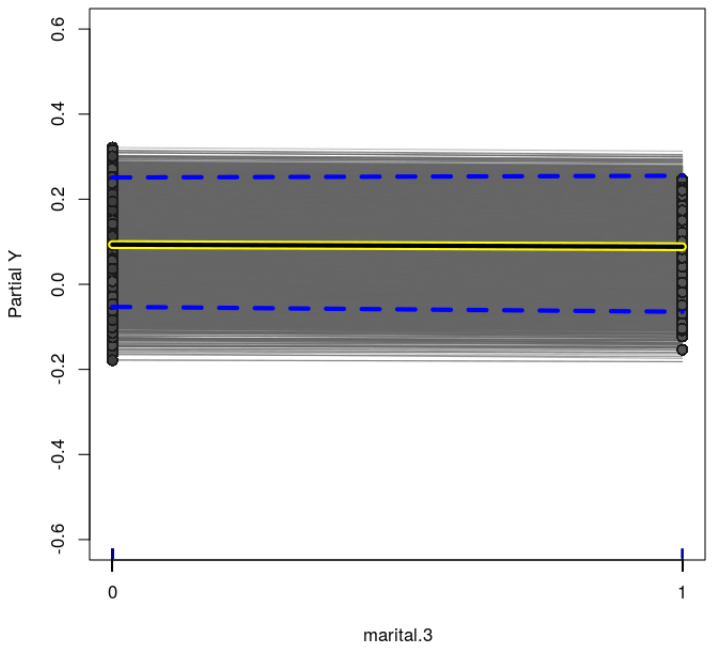

Figure E.22: *BCF model (Real Data Analysis) - Variable marital.3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

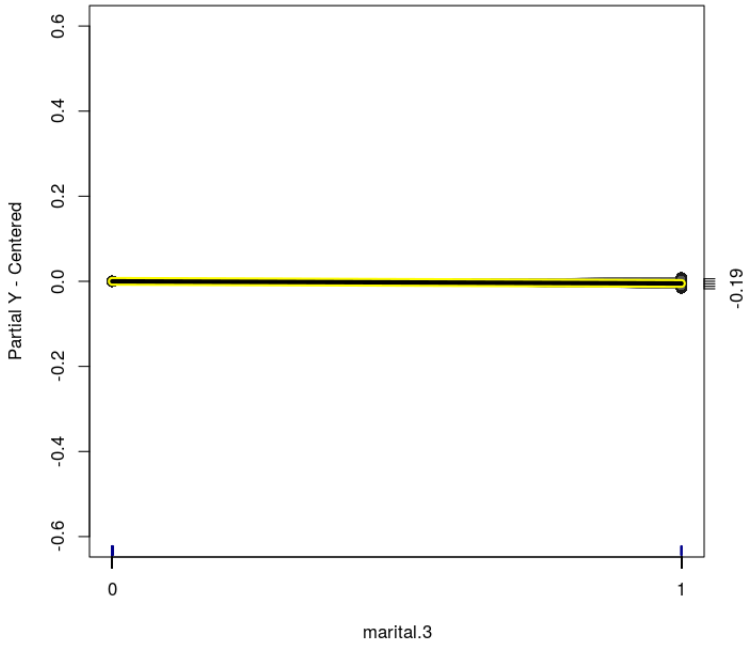

Figure E.23: *BCF model (Real Data Analysis) - Variable marital.3 - centered-ICE Plot for the treatment effect.*

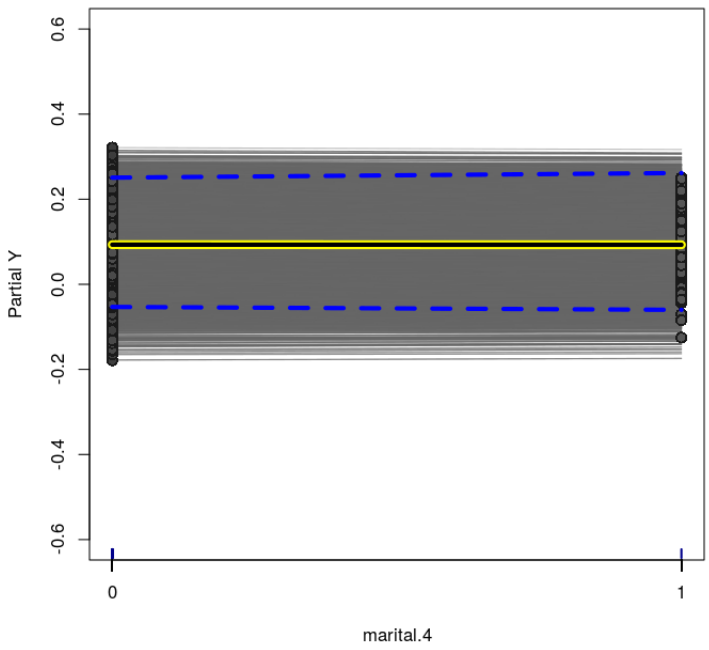

Figure E.24: *BCF model (Real Data Analysis) - Variable marital.4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

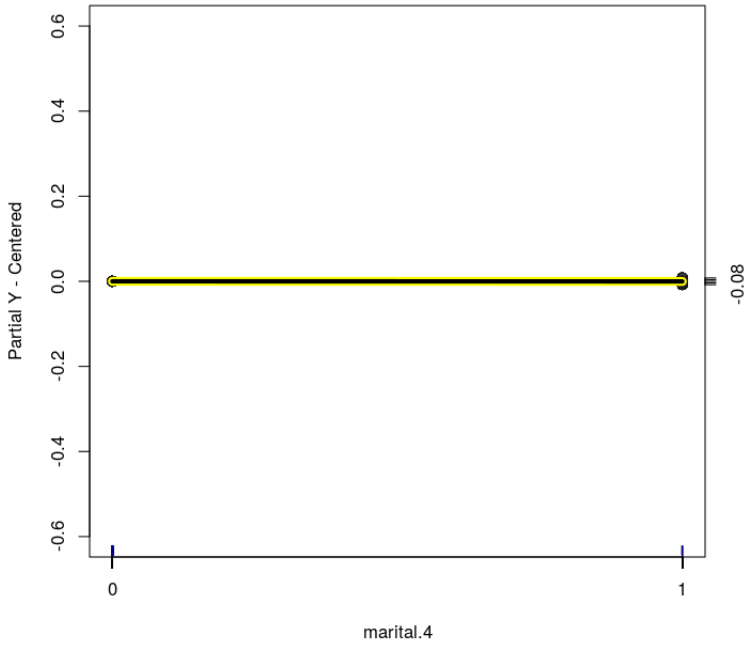

Figure E.25: *BCF model (Real Data Analysis) - Variable marital.4 - centered-ICE Plot for the treatment effect.*

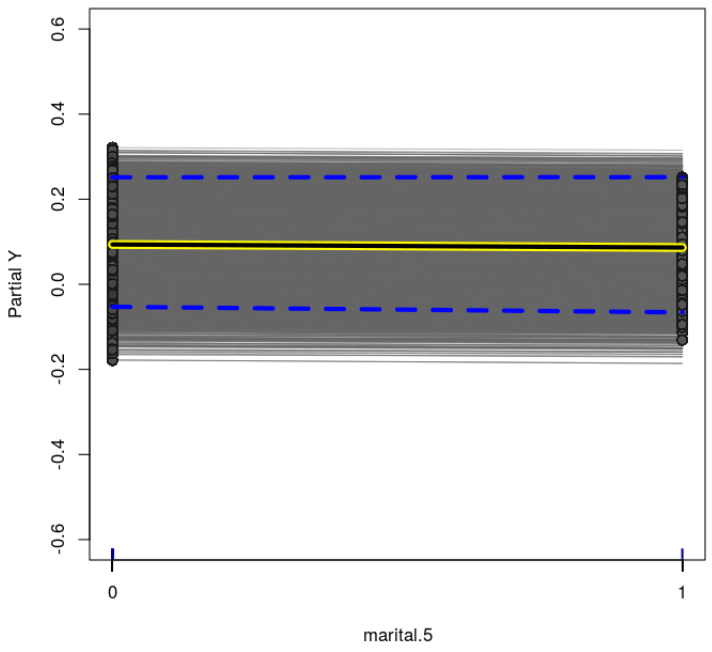

Figure E.26: *BCF model (Real Data Analysis) - Variable marital.5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

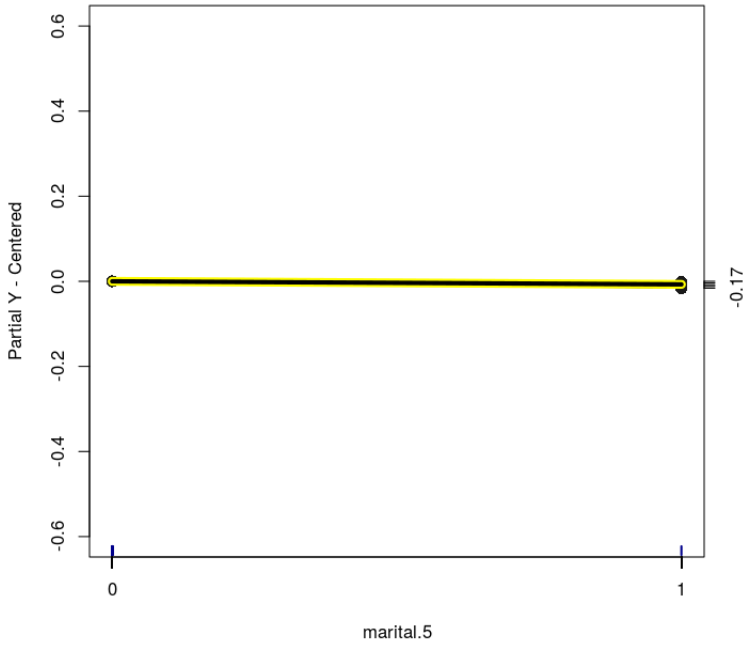

Figure E.27: *BCF model (Real Data Analysis) - Variable marital.*5 *- centered-ICE Plot for the treatment effect.*

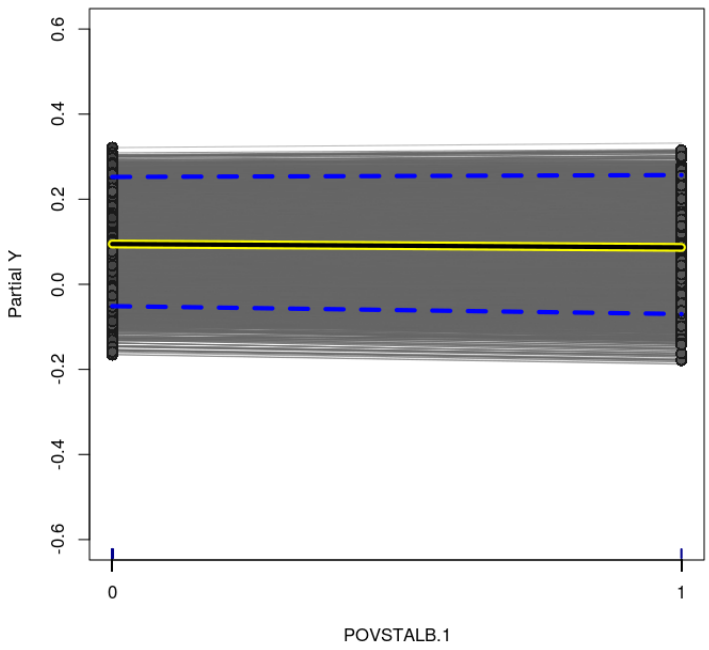

Figure E.28: *BCF model (Real Data Analysis) - Variable POV STALB.*1 *- ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

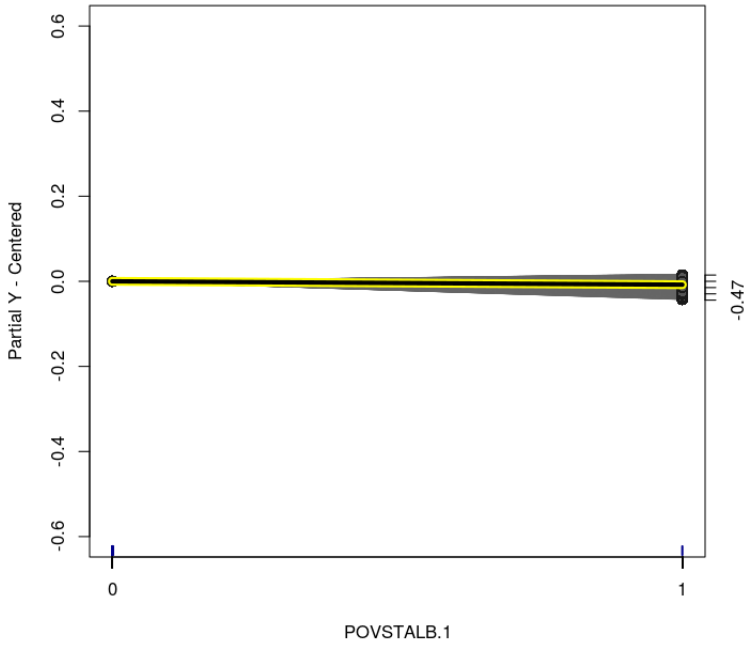

Figure E.29: *BCF model (Real Data Analysis) - Variable POVSTALB.1 - centered-ICE Plot for the treatment effect.*

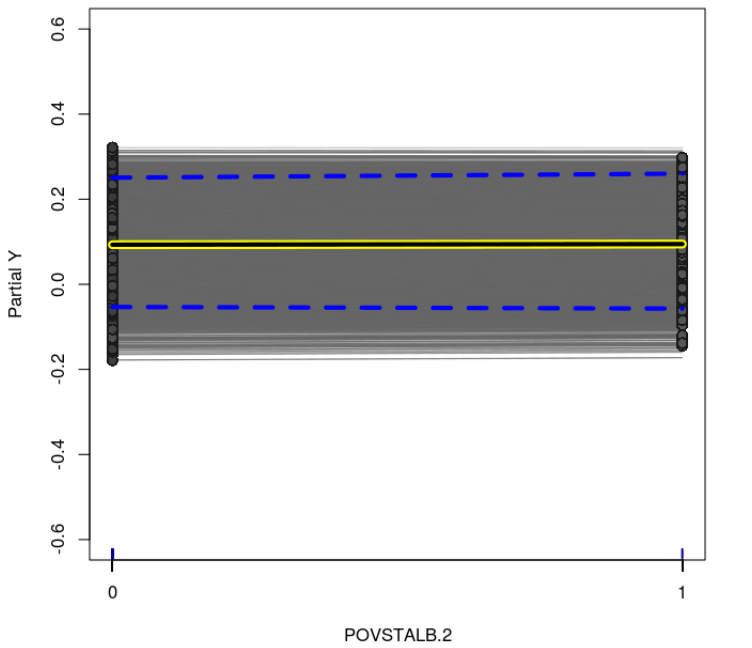

Figure E.30: *BCF model (Real Data Analysis) - Variable POVSTALB.2 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

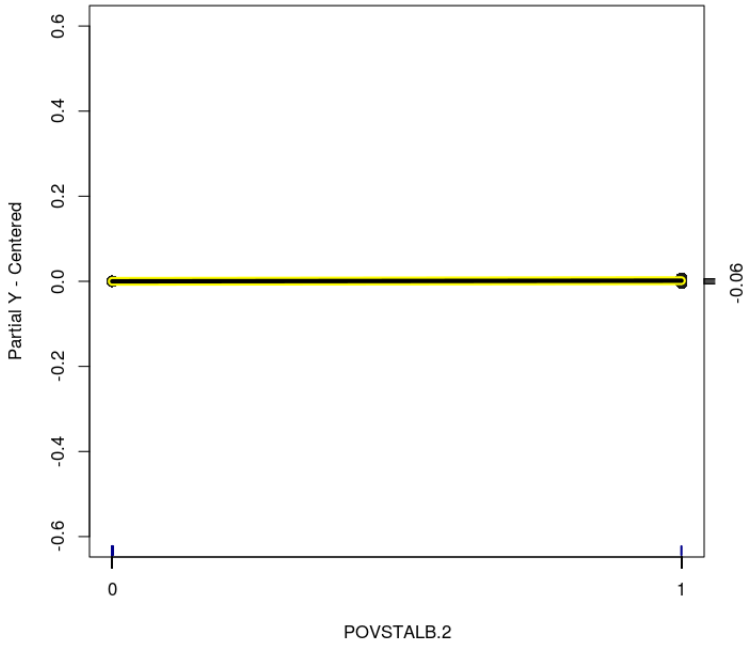

Figure E.31: *BCF model (Real Data Analysis) - Variable POVSTALB.2 - centered-ICE Plot for the treatment effect.*

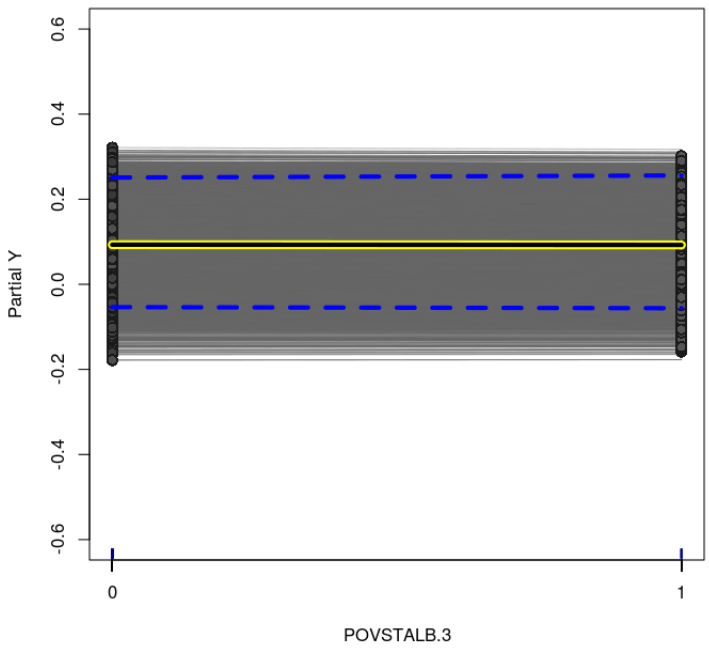

Figure E.32: *BCF model (Real Data Analysis) - Variable POVSTALB.3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

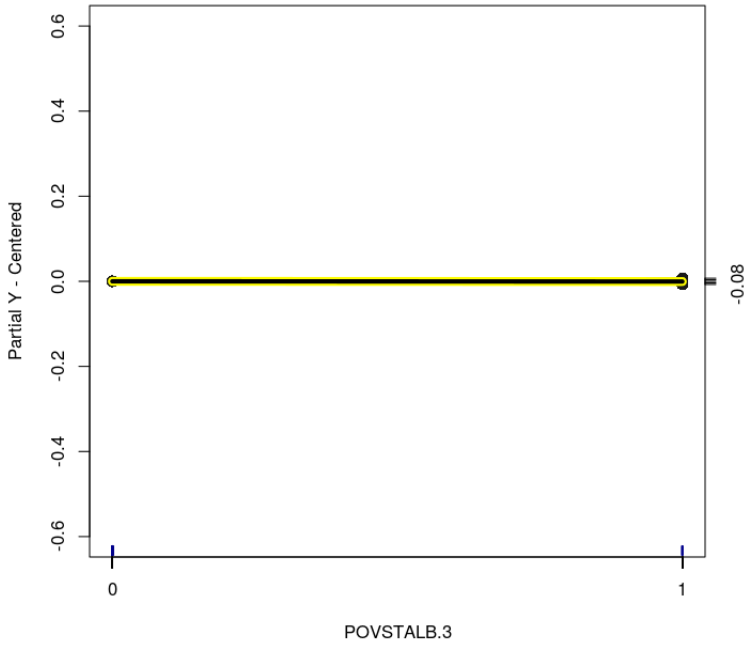

Figure E.33: *BCF model (Real Data Analysis) - Variable POVSTALB.3 - centered-ICE Plot for the treatment effect.*

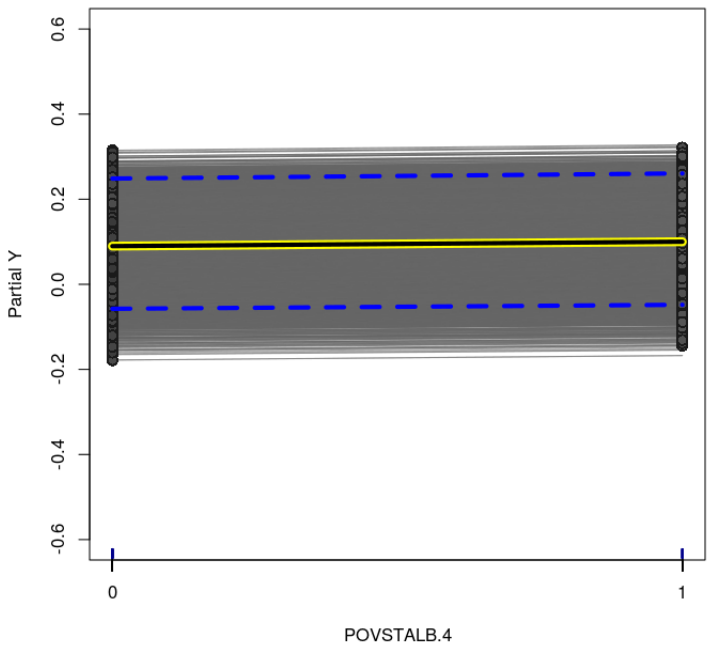

Figure E.34: *BCF model (Real Data Analysis) - Variable POVSTALB.4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

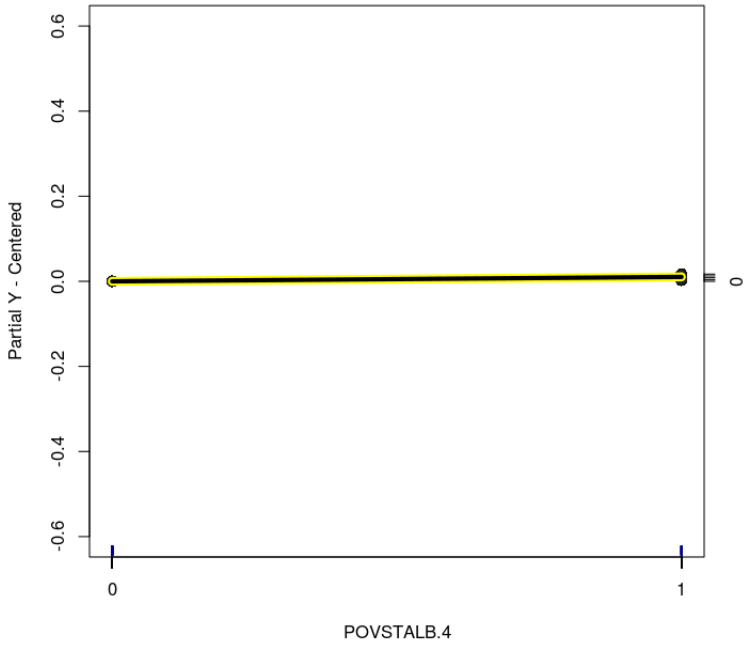

Figure E.35: *BCF model (Real Data Analysis) - Variable POVSTALB.4 - centered-ICE Plot for the treatment effect.*

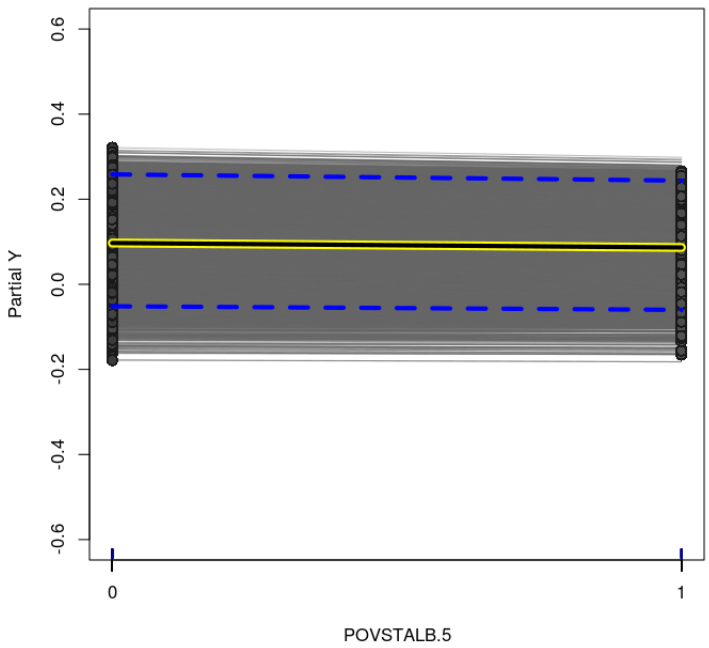

Figure E.36: *BCF model (Real Data Analysis) - Variable POVSTALB.5 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

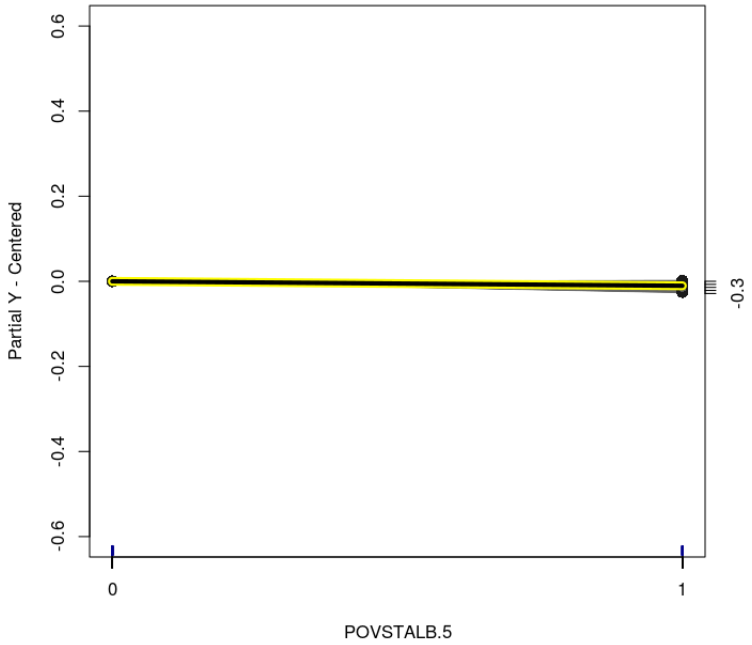

Figure E.37: *BCF model (Real Data Analysis) - Variable POVSTALB.5 - centered-ICE Plot for the treatment effect.*

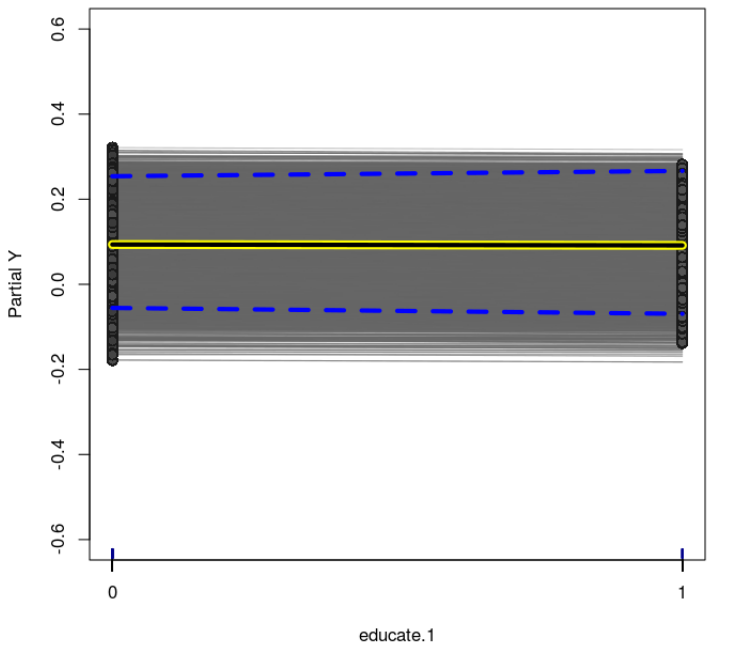

Figure E.38: *BCF model (Real Data Analysis) - Variable educate.1 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

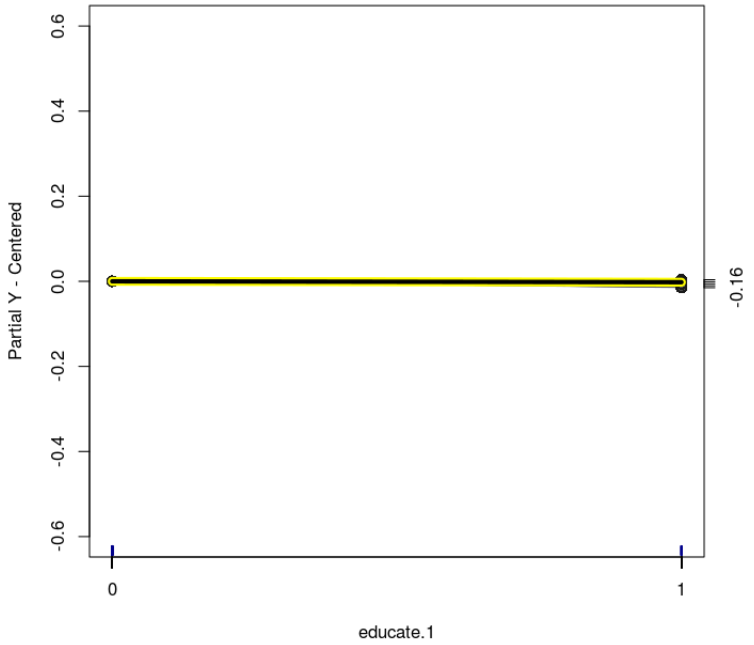

Figure E.39: *BCF model (Real Data Analysis) - Variable educate.*1 *- centered-ICE Plot for the treatment effect.*

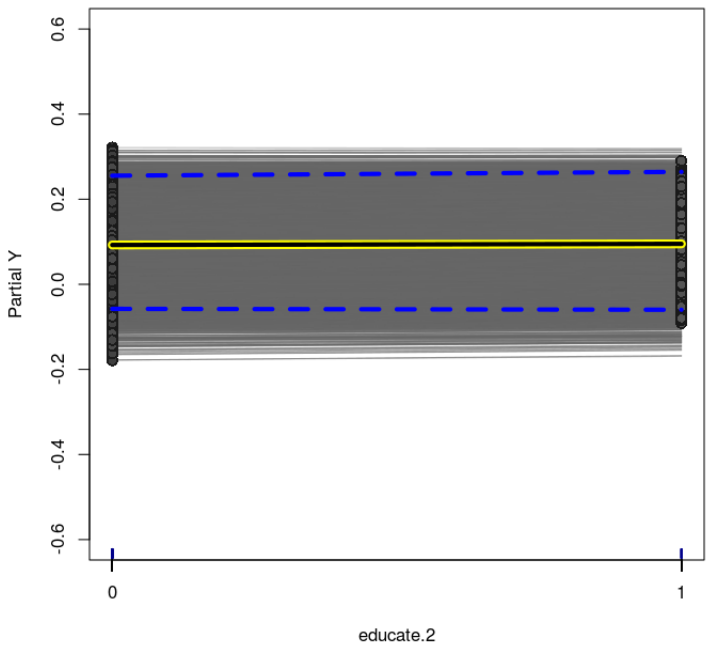

Figure E.40: *BCF model (Real Data Analysis) - Variable educate.*2 *- ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

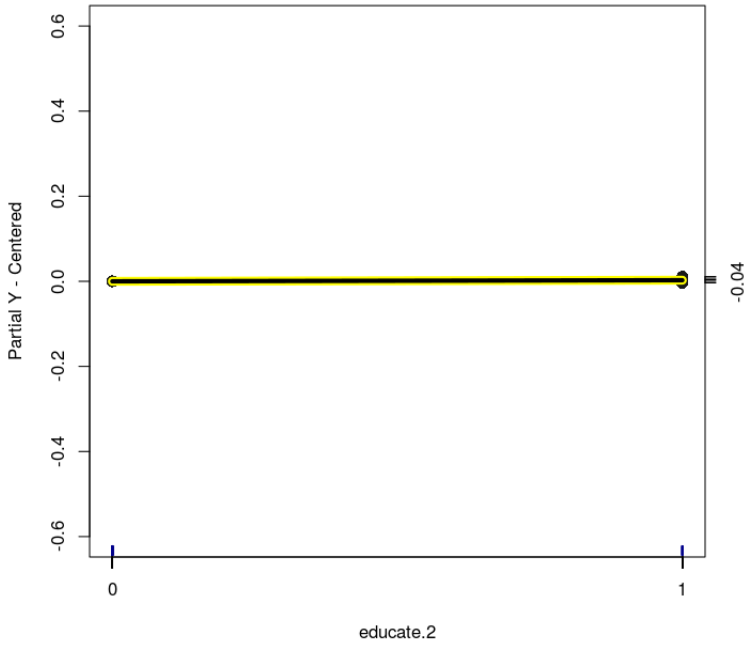

Figure E.41: *BCF model (Real Data Analysis) - Variable educate.*2 *- centered-ICE Plot for the treatment effect.*

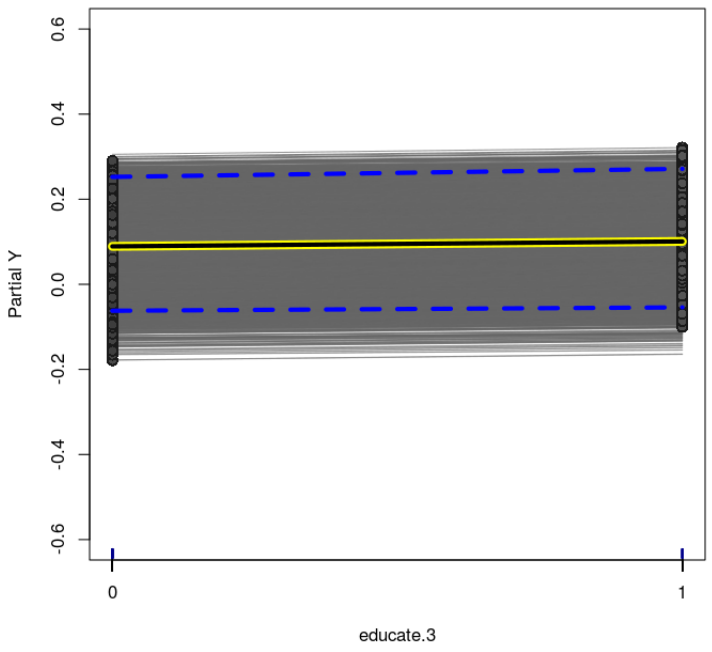

Figure E.42: *BCF model (Real Data Analysis) - Variable educate.*3 *- ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

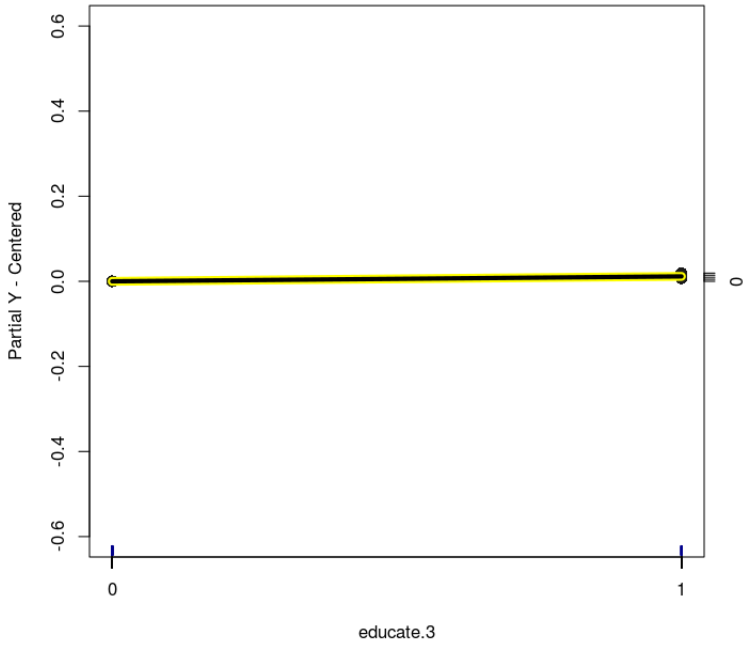

Figure E.43: *BCF model (Real Data Analysis) - Variable educate.*3 *- centered-ICE Plot for the treatment effect.*

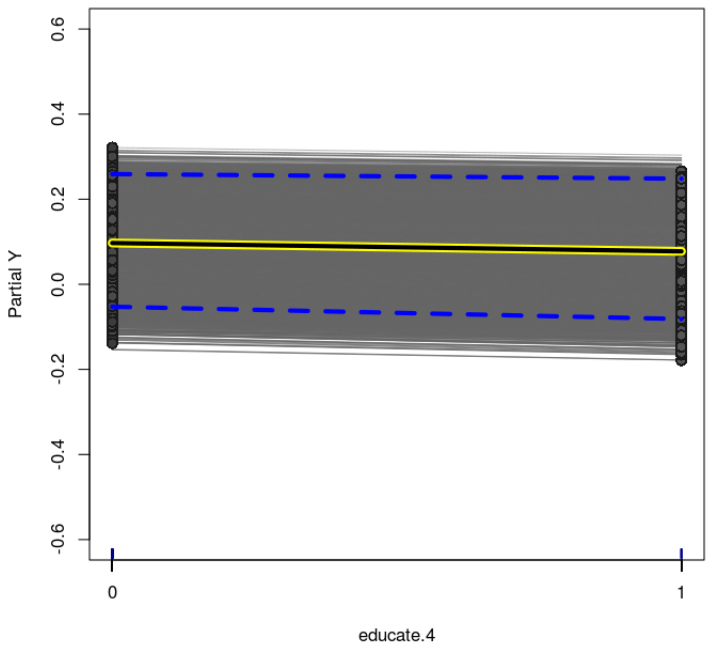

Figure E.44: *BCF model (Real Data Analysis) - Variable educate.*4 *- ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

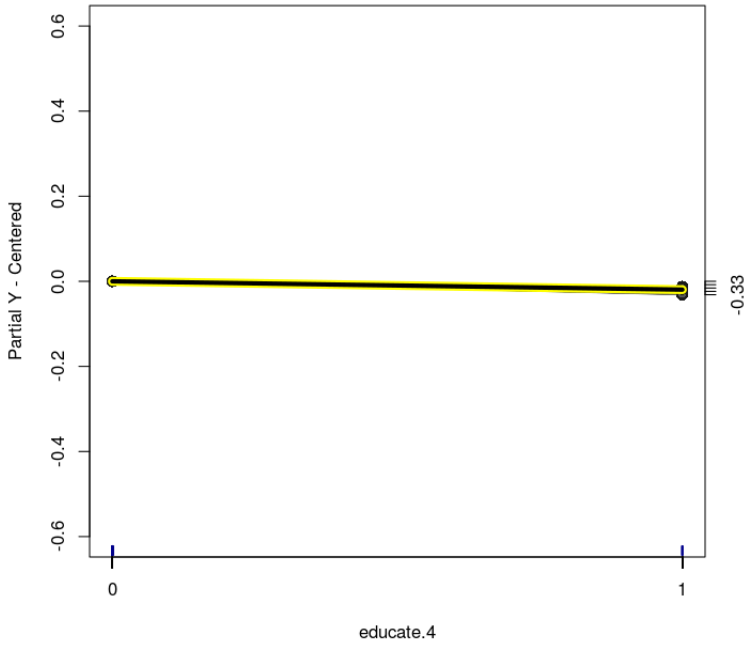

Figure E.45: *BCF model (Real Data Analysis) - Variable educate.*4 *- centered-ICE Plot for the treatment effect.*

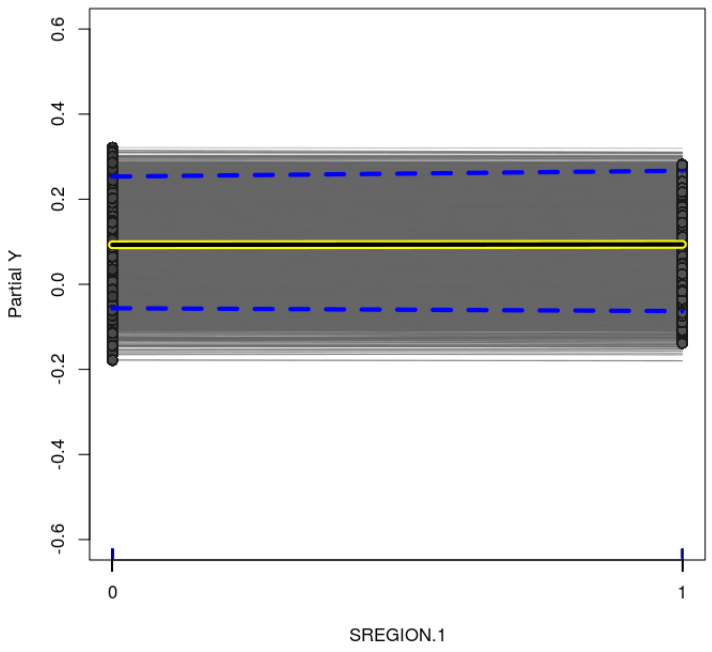

Figure E.46: *BCF model (Real Data Analysis) - Variable SREGION.*1 *- ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

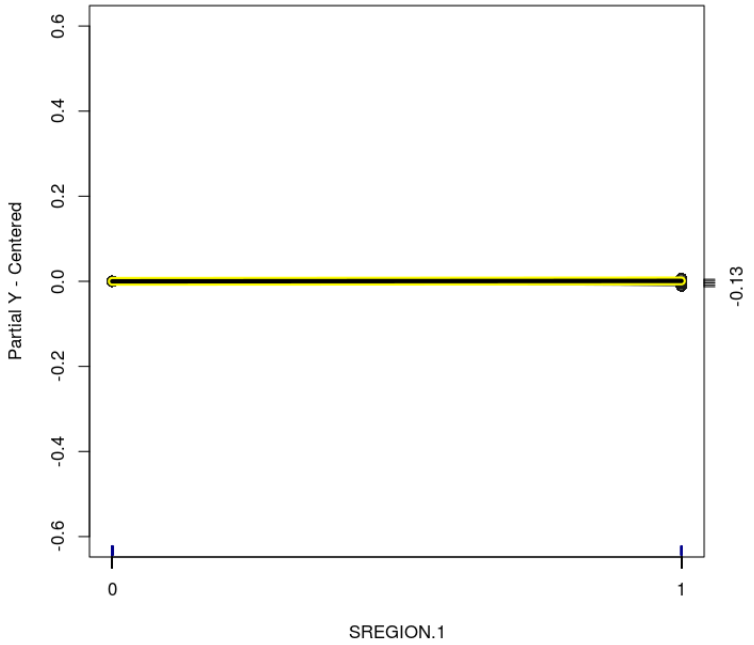

Figure E.47: *BCF model (Real Data Analysis) - Variable SREGION.*1 *- centered-ICE Plot for the treatment effect.*

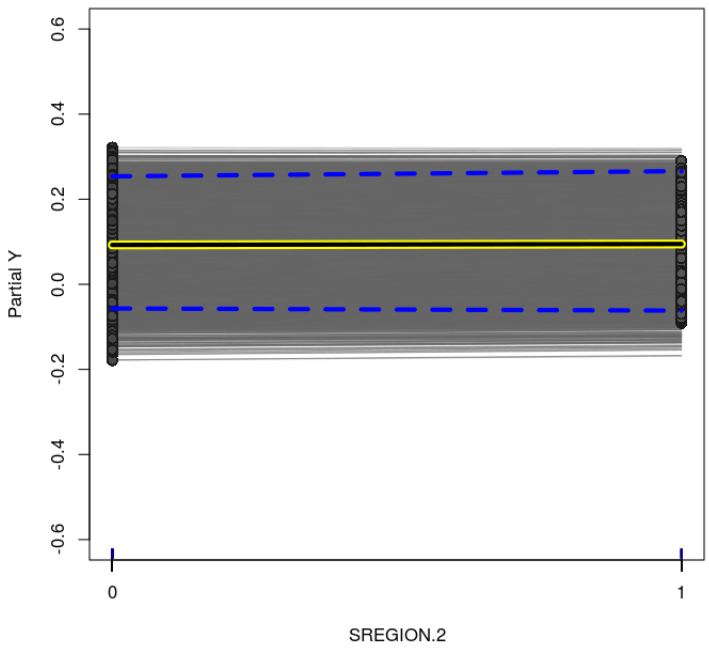

Figure E.48: *BCF model (Real Data Analysis) - Variable SREGION.*2 *- ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

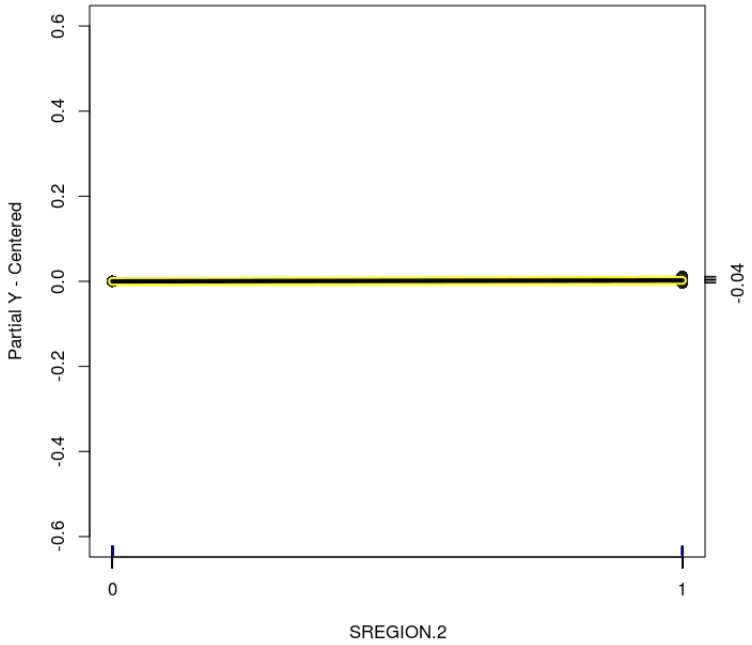

Figure E.49: *BCF model (Real Data Analysis) - Variable SREGION.2 - centered-ICE Plot for the treatment effect.*

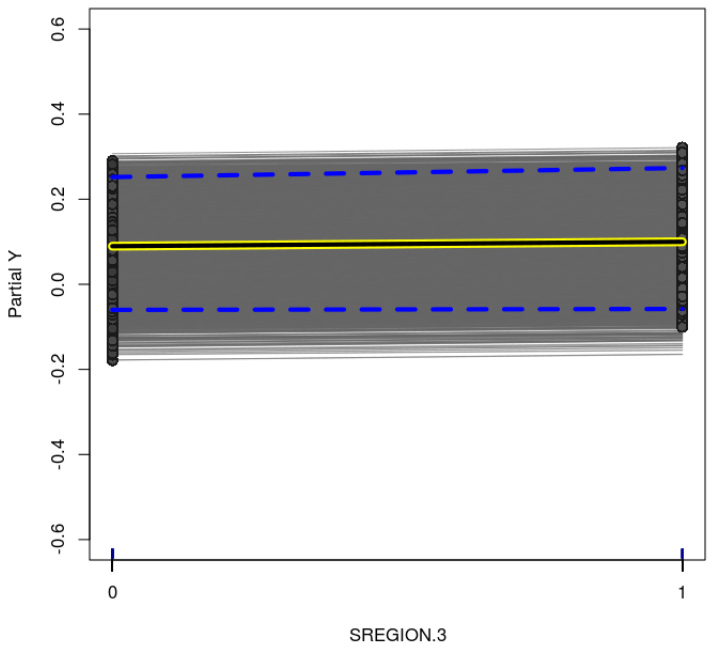

Figure E.50: *BCF model (Real Data Analysis) - Variable SREGION.3 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

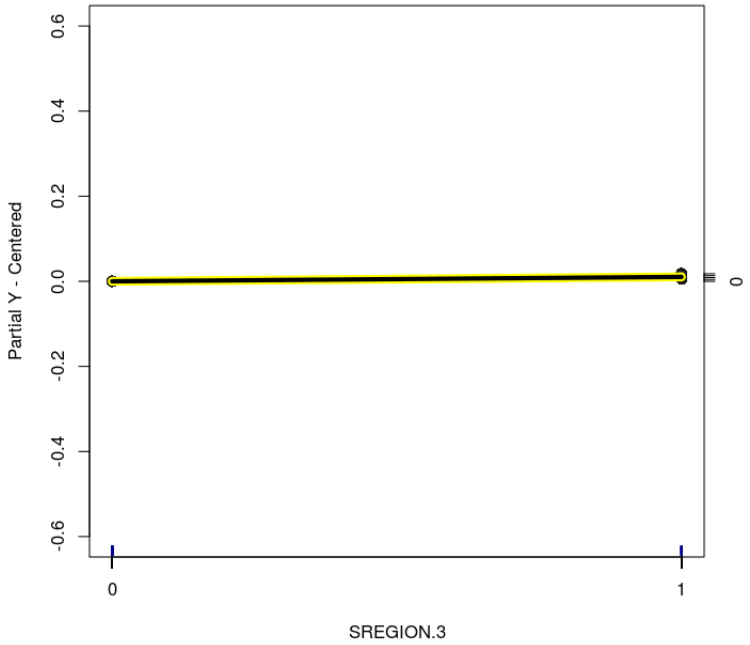

Figure E.51: *BCF model (Real Data Analysis) - Variable SREGION.3 - centered-ICE Plot for the treatment effect.*

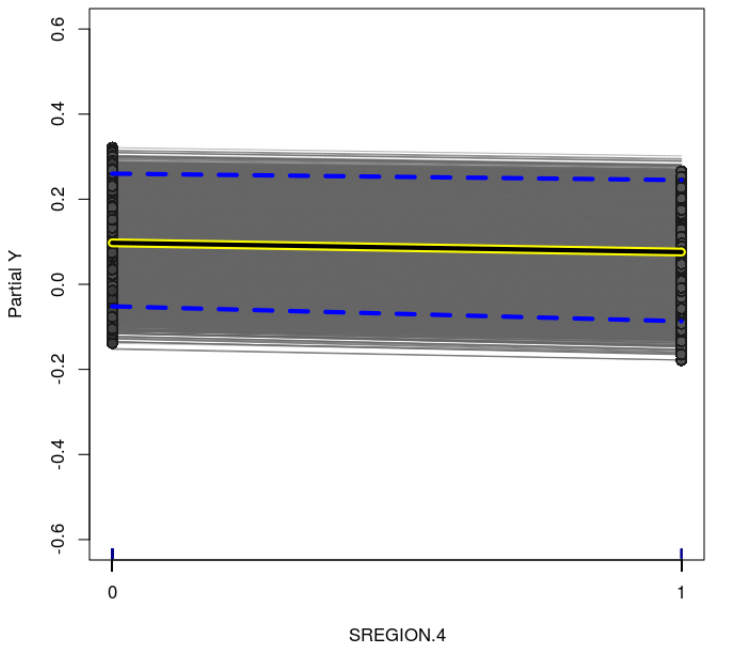

Figure E.52: *BCF model (Real Data Analysis) - Variable SREGION.4 - ICE Plot for the treatment effect. Dashed lines are the 95% credible interval for the estimated PDP.*

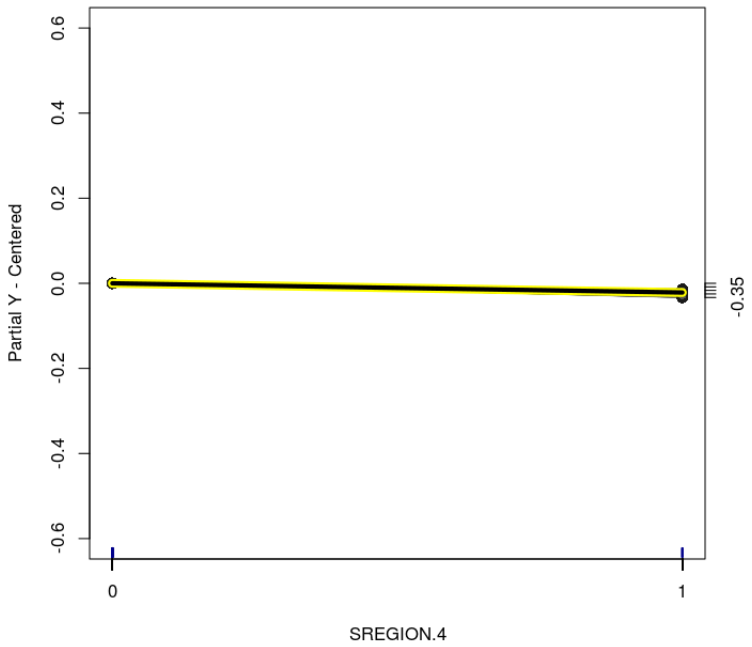

Figure E.53: *BCF model (Real Data Analysis) - Variable SREGION.4 - centered-ICE Plot for the treatment effect.*

Bibliography

Albert and Chib (1993) James H. Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679. From page 16

Barbieri and Berger (2004) Maria Maddalena Barbieri and James O. Berger. Optimal Predictive Model Selection. *The Annals of Statistics*, 32(3):870–897. doi: 10.1214/009053604000000238. From page 30

Bleich *et al.* **(2014)** Justin Bleich, Adam Kapelner, Edward I. George and Shane T. Jensen. Variable Selection for BART: An Application to Gene Regulation. *The Annals of Applied Statistics*, 8(3):1750–1781. doi: 10.1214/14-AOAS755. From page 52

Breiman *et al.* **(1984)** Leo Breiman, Jerome H. Friedman, Richard A. Olshen and Charles J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA. From page 6

Chipman *et al.* **(1998)** Hugh A. Chipman, Edward I. George and Robert E. McCulloch. Bayesian CART Model Search. *Journal of the American Statistical Association*, 93(443):935–960. From page 7, 8, 9, 10, 12, 13, 14, 79

Chipman *et al.* **(2010)** Hugh A. Chipman, Edward I. George and Robert E. McCulloch. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4(1):266–298. doi: 10.1214/09-AOAS285. From page 1, 11, 13, 14, 15, 16, 27, 31, 45

Denison *et al.* **(1998)** David G. T. Denison, Bani K. Mallick and Adrien F. M. Smith. A Bayesian CART Algorithm. *Biometrika*, 85(2):363–377. From page 8

Friedman (2001) Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232. From page 3, 30

Friedman (1991) Jerome H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67. From page 52

Gelman (2006) Andrew Gelman. Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analisys*, 1(3):515–534. doi: 10.1214/06-BA117A. From page 17

Gelman and Hill (2007) Andrew Gelman and Jennifer L. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. From page 21

Geweke (1992) J. Geweke. *Bayesian Statistics. Chapter: Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments*. Clarendon Press, 4 ed. From page 16, 47, 53

Goldstein *et al.* **(2015)** Alex Goldstein, Adam Kapelner, Justin Bleich and Emil Pitkin. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65. doi: 10.1080/10618600.2014.907095. From page 3, 27, 31, 32, 45

Green and Kern (2012) Donald P. Green and Holger L. Kern. Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees. *Public Opinion Quarterly*, 76(3):491–511. From page 31

Hahn *et al.* **(2017)** P. Richard Hahn, Jared S. Murray and Carlos Carvalho. *bcf: Bayesian Causal Forests*, 2017. R package version 1.0.1. From page 45

Hahn *et al.* **(2018a)** P. Richard Hahn, Carlos M Carvalho, David Puelz and Jingyu He. Regularization and Confounding in Linear Regression for Treatment Effect Estimation. *Bayesian Analysis*, 13(1):163–182. doi: 10.1214/16-BA1044. From page 3, 27, 29

Hahn *et al.* **(2018b)** P. Richard Hahn, Jared S. Murray and Carlos M. Carvalho. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects. *arXiv e-print, arXiv: 1706.09523v2*. From page 3, 16, 17, 27, 28, 29, 34, 45, 46, 64, 66, 72, 77, 78

Hastie and Tibshirani (2000) Trevor Hastie and Robert Tibshirani. Bayesian Backfitting. *Statistical Science*, 15(3):196–223. From page 13

Hastie *et al.* **(2001)** Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA. From page 5

Hill (2011) Jennifer L. Hill. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240. doi: 10.1198/jcgs.2010.08162. From page 3, 14, 17, 27, 28, 29, 45, 46, 77

Imai and van Dyk (2004) Kosuke Imai and David A. van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866. From page 66

Imbens (2004) Guido W. Imbens. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics*, 86(1):4–29. From page 28

Johnson *et al.* **(2003)** Elizabeth Johnson, Francesca Dominici, Michael Griswold and Scott L. Zeger. Disease cases and their medical costs attributable to smoking: An analysis of the national medical expenditure survey. *Journal of Econometrics*, 1(112):135–151. From page 64

Kapelner and Bleich (2016) Adam Kapelner and Justin Bleich. bartMachine: Machine Learning with Bayesian Additive Regression Trees. *Journal of Statistical Software*, 70(4):1–40. doi: 10.18637/jss.v070.i04. From page 9, 14

Linero (2017) Antonio R. Linero. A Review of Tree-Based Bayesian Methods. *Communications for Statistical Applications and Methods*, 24(6):543–559. doi: 10.29220/CSAM.2017.24.6.543. From page 84

Linero (2018) Antonio R. Linero. Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection. *Journal of the American Statistical Association*. doi: 10.1080/01621459.2016.1264957. From page 3, 27, 29, 30, 52

McCulloch *et al.* **(2019)** Robert McCulloch, Rodney Sparapani, Robert Gramacy, Charles Spanbauer and Matthew Pratola. *BART: Bayesian Additive Regression Trees*, 2019. URL https://CRAN.R-project.org/package=BART. R package version 2.2. From page 45

Neyman (1923) Jerzy Neyman. On the application of probability theory to agricultural experiments: Essay on principles. section 9. english translation (1990) by d. m. dabrowska and t. p. speed. *Statistical Science*, 5(4):465–480. From page 19

Pearl (2000) Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press. From page 19

Pearl (2009) Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 ed. From page 25

Pratola *et al.* **(2017)** M. T. Pratola, H. A. Chipman, E. I. George and R. E. McCulloch. Heteroscedastic BART Using Multiplicative Regression Trees. *arXiv e-print, arXiv:1709.07542*. From page 78

Pratola (2016) Matthew T. Pratola. Efficient Metropolis-Hastings Proposal Mechanisms for Bayesian Regression Tree Models. *Bayesian Analysis*, 11(3):885–911. doi: 10.1214/16-BA999. From page 14, 78

R Core Team (2017) R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/. From page 45

Rosenbaum and Rubin (1983) Paul R. Rosenbaum and Donald B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55. From page 21, 22, 23, 24, 27, 28

Ročková and van der Pas (2019) V. Ročková and S. van der Pas. Posterior Concentration for Baysian Regression Trees and Forests. *arXiv e-print, arXiv:1708.08735v5*. From page 14, 43

Rubin (1974) Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701. From page 19

Sparapani *et al.* **(2019)** Rodney Sparapani, Charles Spanbauer and Robert McCulloch. *Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: the BART R Package*, 2019. URL https://CRAN.R-project.org/package=BART. From page 14, 16, 40, 47, 53

van der Pas and Ročková (2017) S. van der Pas and V. Ročková. Bayesian Dyadic Trees and Histograms for Regression. *arXiv e-print, arXiv:1708.00078*. From page 14

Zigler and Dominici (2014) Corwin Matthew Zigler and Francesca Dominici. Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model Average Causal Effects. *Journal of the American Statistical Association*, 109(505):95–107. doi: 10.1080/01621459.2013.869498. From page 45