EN.625.725 Final Paper: Bayesian Causal Forests for Broken Hearts

Lauren Kimpel

12 December 2022

0.1   Executive Summary

For years, medical literature has supported a link between genetics, age, and heart disease and stroke risk. There is also evidence which suggests that biological sex plays an important role in heart-related disease outcomes [11]. Whether this difference between the sexes is causative is a natural question. The goal for the overall experimentation will be to predict heart-based anomalies based on biological sex. We will also observe the performance of a Bayesian Regression Tree (BART) model on heterogeneous treatment effects, as detailed in [1]. This strategy employs techniques from causal data analysis, regression testing, and Bayesian inference, so their exploration will be our primary focus in this document. Also included in the collection of methods are the tools employed in the initial data-testing phase, including sample estimator calculation, covariate hypothesis testing and data plotting.

   The organization of this document is as follows: the first and second sections of this document intend to provide a detailed description of the goals and objectives of this project, as well as introduce concepts to be expanded upon later. The third section details the formal methods used in performing our statistical analysis, which range from initial frequentist observations to BART and BCF models. The fourth section aims to discuss the objective results obtained by studying heart disease data under the conditions detailed in section 3. Finally, the fifth and sixth sections are allocated for conclusions and appendices.

0.2   Project Description

Participants & Data

Our sample consists of $n = 1025$ individuals under 14 categories. Of the 1025 participants, 713 (roughly 70 percent) are men and 312 (around 30 percent) are women. This data, collected in 1988, originates from four databases: Cleveland, Hungary, Switzerland, and Long Beach V. In particular, we are interested in evaluating the role of biological sex as a predictor of heart disease. Hence, from a regression standpoint, biological sex is the covariate X, and the presence of heart disease is the designated response Z.
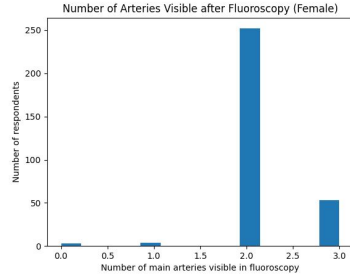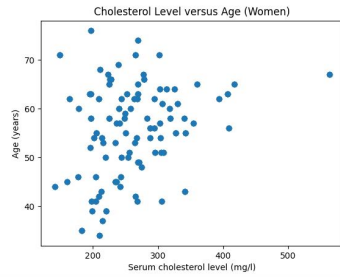
(a) Female Results
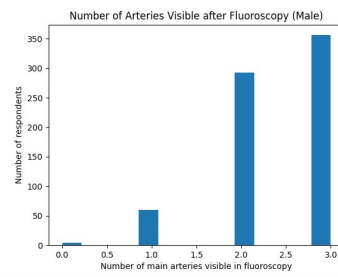
(b) Male Results

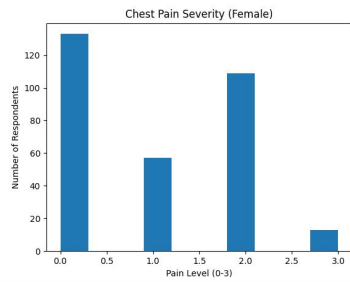Figure 1: Visibility of Main Arteries Between Male and Female Patients
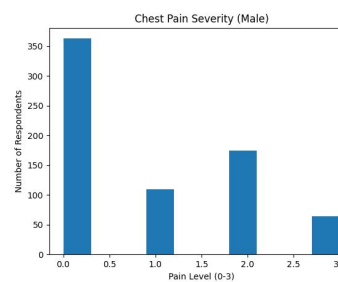

(a) Female Results

(b) Male Results

Figure 2: Age Versus Cholesterol Levels (mg/l) Between Male and Female Patients

(a) Female Results

(b) Male Results

Figure 3: Reported Chest Pain (Scale Between 0-3)

0.2.1 Preliminary Analysis

Initial Analysis, Covariates & Outcomes

To justify our suspicions, we first do some plotting. This will help us develop hypotheses for the relationships between variables in our dataset. In particular, we have chosen to segregate the visualizations by gender.

Below are a few of the ways we might test for heart disease [5]:

1. Exercise Tests

2. Fluoroscopy Tests

3. Cholesterol Level Exams

4. Electrocardiogram (EKG) Tests

Given these tests, the diagnostic criteria for heart disease includes (but is not limited to) lack of presence of arteries in fluoroscopy, high blood pressure, high cholesterol, and poor electrical response. Intuitively, there should be a trend between poor performance in these medical exams and the presence of heart disease. Simply from the sample means in Table 1, an observer might notice that women consistently present poorer scores than men, despite being fewer in number in the dataset. These **heterogeneous effects**, with the assistance of a decent learning model, can help predict such a discrepancy.

Observations of Sample Mean, Variance, and Standard Error & Categories For Testing

For simplicity, we will include those categories listed in Table 1 (including more if necessary.) By observation, it seems that a larger number of women than men in the dataset have outcomes which indicate heart disease, but, due to the sizeable difference in sample size between men and women, we cannot know for certain the nature of this difference. which makes this dataset a good candidate for a study of causal inference.

We can initially make a surface-level comparison of the significance between each estimated parameter. This can be done using BART techniques; see Section 0.3.1 for more details on how this will be incorporated into our analysis.

The examples in Figure 1 and Figure 2 illustrate a level of ambiguity when comparing each of the heterogeneous risk factors between men and women. We will perform multiple bayesian regression tree techniques over a subset

Sample Means by Gender (m=713, f=312)		
	Male	Female
Pain Level	0.92	1.0
Blood Pressure	130.7	133.7
Number Visible Arteries	2.4	2.1
Cholesterol Levels (mg/l)	239.3	261.5
Heart Disease True (% of Participants)	42.1	72.4
Age	53.7	55.8

Table 1: Sample means between men and women in the heart disease dataset.

of the 14 columns of data (`age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, target`).

Causal Inference Analysis Tools

To examine random variables for causality, we refer to **treatment** as an effect. Typically (but not always), this takes on a binary meaning; for example, drinking 12 cups of coffee per day is a treatment. If a treatment variable X is associated with drinking 12 cups of coffee per day, we can denote $X = 1$ as "treated" and $X = 0$ as "not treated."

Define C_0 to be the outcome if a subject is not treated, and define C_1 to be the outcome if the subject is treated.

Then the **average causal effect** is said to be the consistent estimator

$$\theta = \mathbb{E}(C_1) - \mathbb{E}(C_0)$$

. For our purposes, we will consider as treatments whether the subject is a man or a women. The outcomes of the treatment are simply the data points that occur for men and women, respectively. Let $X = 0$ be representative of women and $X = 1$ representative of men. Then according to [3], the conditional causal effect for relative to men is described as

$$\theta_1(C_1) = \mathbb{E}(C_1|X = 1) - \mathbb{E}(C_0|X = 1)$$

and likewise, θ_0 is defined similarly for women. Thus, C_0 is the outcome if the subject is a woman, and similarly, C_1 is the outcome if the subject is a man.

However, conditional causal effect is not to be confused with **association**, which is defined by

$$\alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0),$$

where Y is the **consistency relationship**

$$Y = \begin{cases} C_0 & X = 0 \\ C_1 & X = 1 \end{cases}$$

The following sections are dedicated to each of the causative tools we will use to build and analyze our model.

Random Variable Definition & Consistency Relationship

For this project, let $C* = (X, Y, C_0, C_1)$ be a vector whose representations are described above. That is, X is a binary treatment variable representative of biological sex; Y is its consistency relationship. C_0 and C_1 are the outcomes for female and male subjects, respectively, in a particular heart disease category belonging to those listed in Table 1.

Conditional Causal Effect Estimator Calculation

Consider a subset of the heart disease data relative to the outcome of heart disease in men and women.

Recall the sample means for both $X = 0$ and $X = 1$, as described in Table 1, which are $\mathbb{E}(C_0|X = 0) = 42.1$ and $\mathbb{E}(C_1|X = 1) = 72.4$. Then $\theta_1 = 42.1$ and $\theta_0 = 72.4$, since both C_1 and C_0 are counterfactuals - that is, if $X = 1$, we don't observe C_0, and vice versa [3].

For our purposes, we'll use expected values for Age, Pain Level, Blood Pressure, Artery Visibility, and Cholesterol Level for both men and women to describe the conditional causal parameters θ_1 and θ_0.

The model will then validate these parameters against the true outcomes (heart disease or no heart disease.)

0.3 Methods

0.3.1 Bayesian Causal Forests

An **inference** over a data point x is defined to be

$$Y = f(x) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2). \qquad\qquad [9]$$

A Bayesian Additive Regression Tree (BART) model wishes to approximate $f(x)$ by obtaining $\mathbb{E}(Y|f(x))$, where $f(x)$ is expressed as the sum of bayesian regression trees g_i over x. In other words, $f(x) \approx \sum_{i=1}^{N} g_i(x)$. This expression for f is also the prior distribution for the BART model. Though the classic BART model is historically good for inference [1][7], this inference does not necessarily imply causality.

Causal inference is the study of bidirectionality in causality; if X, Y are random variables, and X impacts Y, causal inference tools wish to know whether Y also impacts X [3].

This project is devoted to applying a proposed improvement on the BART-based model, known as Bayesian Causal Forests (BCF), for the sake of flexible causal inference. In particular, we want to examine whether biological sex is not only a prediction, but a *cause* of aggregated heart-related issues.

The main difference between BART and BCF is that the latter is more concerned with reparameterizing its priors to accomodate additional conditional causal effect statistics. This impacts where its decision trees must split, and provides greater flexibility (and resistance to regression phenomena such as RIC) when examining heterogeneous effects.

In the first part of this section, we will develop the basic tools necessary for producing a discussion of causality. We will then demonstrate how they will be used in our experimentation. Finally, we will examine the formation of priors for the BART and BCF models, and enumerate the regression techniques that will be used in conjunction with our application of the causal forest model.

Modeling Overview

We wish to develop and compare 2 models for this project. We will use the `bartMachine` package in R to perform a default BART analysis of the data (estimating `target`), in addition to propensity score calculation. Finally, the

bulk of the BCF simulation will be done using the `BCF` package. We will compare the strength of inference in both models. For simplicity, we will employ each with default settings (the explicit prior parameterizations for `BCF` are mentioned in later sections.) We will estimate `target` under the constraint of biological sex, in addition to the various risk factors (heterogeneous effects) associated with heart disease, which belong to the categories listed in Table 1.

Development of the BCF prior

In the BART methodology, a prior $f(x)$ is chosen by approximating a sum of binary regression trees g_i. The trees themselves contain internal decision nodes, whose structure is defined by splitting rules. The splitting rules for a binary regression tree are determined by $f(x)$. We can express a prior as the additive regression forest $f(x) = \sum_{i=1}^{N} g_i(x)$, where $g_i(x)$ maps to some scalar. In ordinary Bayesian analysis, choice of prior is designed to be an uninformed event, meaning that we will select a prior distribution $f(\theta)$ and statistical model $f(x|\theta)$ over some parameter θ with little insight into the true distribution. After viewing our results, we then update our parameters.

Per the initial introduction of BART into statistical literature, probability that a node at depth h splits is precisely $\eta(1 + h)^{-\beta}$, where $\eta \in (0, 1)$ and $\beta \in [0, \infty)$ [5].

For BART, quality of prior does matter. It is also assumed that the structure of each decision tree g_i and its leaves m_i are independent. Hence, prior specification can be segregated into parts if desired. For our purposes, we will be going with the naive approach, avoiding this option.

Per [5], large, deep trees take $\eta = 0.95$ and $\beta = 2$. N^{th} leaves are distributed according to $\mathcal{N}(0, \sigma^2)$, with $\sigma = \frac{\sigma_0}{\sqrt{N}}$.

Implications for Causal Inference

For causal analysis, however, the BART prior selection criteria do not necessarily account for causal priors; e.g., the inference of $\theta(x) = f(x|1) - f(x|0)$, or the mean conditional causal effect described in previous sections.

We will be naively following the BCF prior selection criteria given by [1], which is based on **propensity scoring**, or calculation of the probability $P(X = a|x)$, where $a \in \{0, 1\}$. This is not the same as the conditional causal effect estimator $\theta_a(x)$ described above. Set $\eta = 0.25$ and $\beta = 3$. In addition,

we will reparamaterize the model to fit into a $\mathcal{N}(\mu + \theta_a(x), \sigma^2)$ distribution, as opposed to $\mathcal{N}(\mu, \sigma^2)$.

Additional Hypothesis Testing

Let H_0 be the statement that heart-disease risk is not significantly motivated by biological sex. Let H_1 be the alternate hypothesis, that biological sex has causes inflated heart disease risk.

"Heart disease risk" here can be thought of an aggregation of the following results:

1. significantly higher blood pressure measurements

2. fewer visible arteries in fluoroscopy

3. higher cholesterol levels

4. greater chest pain measurements

We can test multiple covariates for non-causal significance against our response Z. Generate k BART models, permuting differently over each covariate vector k times, and note measure-of-fit on each run.

Let β be the number of pseudo r-squared values such that $\beta > \hat{\beta}$, where $\hat{\beta}$ is the observed pseudo r-squared value. Then the p-value is given by

$$\frac{\beta}{k+1} \qquad [10]$$

To inform our covariate selection, we will use the sample means of the 4 categories listed above as our parameters for comparison.

We will be using the `cov_importance_test` function from the `bartMachine` library in R, which implements the above analysis technique.

0.3.2 Regression Testing Tools

Unsurprisingly, the *regression* aspect of the BART and BCF methods is what drives the inference and learning of $f(x)$. It is not typical for us to have access to extensive knowledge of our parameters. Therefore, BART and BCF are **nonparametric** learning models, which makes them particularly flexible.

In the BART model, regression is expressed as the output of multiple decision trees over data point estimation given some prior distribution. In a BCF, this idea is still supported.

Fitting

Because we are adding in additional information regarding θ, we must also penalize the BCF regression scheme, else we risk overfitting our data. [1]

To solve this, the BCF improvement over BART **regularizes** the difference between the two conditional expectations in the body of θ. Though advanced fitting options are explored in a variety of causal forest models, we will use the `BCF` and `bartMachine` default settings for simplicity. This is equivalent to the regularization modifications featured above.

Targeted Selection

The regression model proposed in [1] attempts to alleviate bias by incorporating **targeted selection**, or estimation of what a value *would* be in the absence of a treatment effect [1].

In Section 0.2.1, it is noted that our dataset suffers from a discrepancy in sample populations between men and women. Using targeted selection, we can predict effects for women, essentially generating additional new data for testing. Unfortunately, this can lead to overfitting.

The BCF method modifies prior selection to be *covariate-dependent*, so that we can avoid the overfitting problem and further restrict regression (see Section 0.2.3).

Time permitting, we will implement targeted selection using BCF, and see whether this significantly impacts the predictions made using the initial BCF model.

0.3.3 Explicit Implementation Summary

Below is the `bcf` function employed by the R-BCF package, with default parameters.

```
bcf(y, z, x_control, x_moderate = x_control, pihat, nburn, nsim, nthin = 1,
   update_interval = 100, ntree_control = 200, sd_control = 2 * sd(y),
   base_control = 0.95, power_control = 2, ntree_moderate = 50,
   sd_moderate = sd(y), base_moderate = 0.25, power_moderate = 3, nu = 3,
   lambda = NULL, sigq = 0.9, sighat = NULL, include_pi = "control",
   use_muscale = TRUE, use_tauscale = TRUE)
```

We'll be running this particular function using the defaults, in addition to calculations needed for constructing a model.

Below are the calculations and descriptions for each relevant parameter [12].

i z is the 1025-sample treatment variable in consideration: that is, biological sex, taking on values of 0 or 1.

ii $\tau(x_i) = \mathbb{E}(Y_i|x_i, Z = 1) - \mathbb{E}(Y_i|x_i, Z = 0)$ for each individual treatment effect x_i. In the implementation, τ is an $n \times n$ matrix consisting of posterior samples of individual treatment effects.

iii q is the targeted selection criteria.

iv $\mu(x_i) = q + \tau(x_i) \cdot z$. In the implementation, μ is an $n \times n$ matrix consisting of posterior samples of individual treatment effects.

v π is the propensity score estimate. Here, we simply use the default setting.

vi σ is the noise level relative to the expected mean function of Y. Here, $\sigma = \text{sd}(q + \tau * \hat{\pi})$, which is simply twice the estimated standard deviation of $q + \tau$. This raises the "difficulty" of the estimate.

vii $y_i = \mu(x_i, \pi_i) + \tau(x_i, \pi_i)z_i + \epsilon_i$, or the posterior approximation of a response given a treatment variable

viii $\hat{\pi}$ is the optional estimator for the propensity score estimate π.
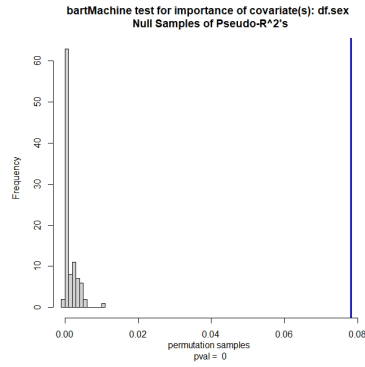
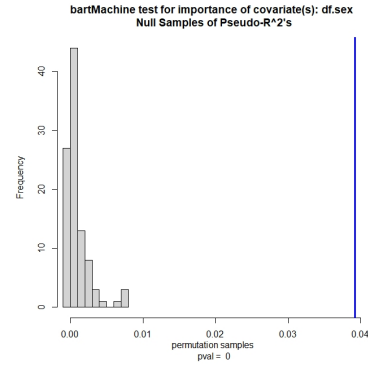## 0.4   Results

0.4.1   BART Hypothesis Testing

We ran `covariance_importance_test` using sex as our binary treatment variable against several different categories of covariate. We obtained several p-values for the following categories, under the usual constraint of $\alpha = 0.05$.

- Presence of heart disease $(p \approx 0.00)$

- Cholesterol levels $(p \approx 0.00)$

- Age $(p = 0.01)$

- Visibility of Arteries $(p \approx 0.00)$

- Chest pain $(p = 0.238)$

- Blood pressure $(p = 0.01)$

Hence, we have reason to believe that this result is statistically significant (with the exception of pain level, the p-value of which suggests statistical *in-significance*.) Naturally, however, causality is not implied. The graphical results of the hypothesis tests are shown in Figure 4.
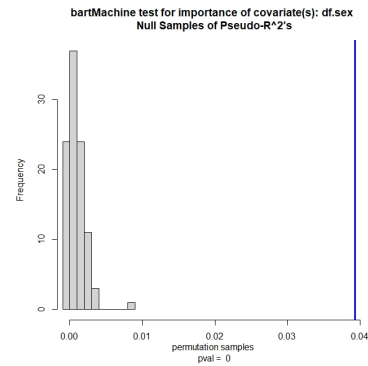
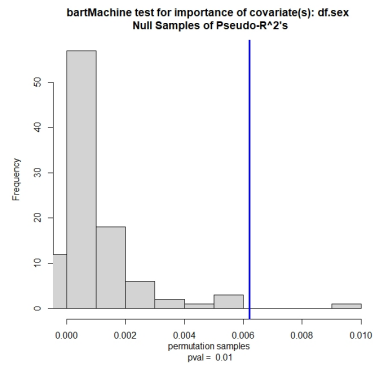(a) sex vs. heart disease diagnosis
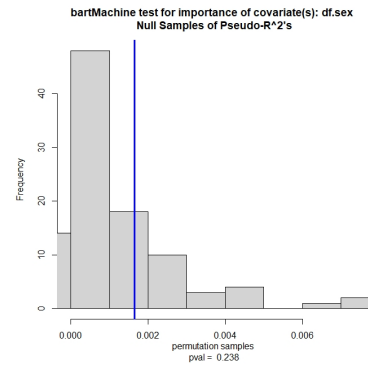
(b) sex vs. cholesterol levels
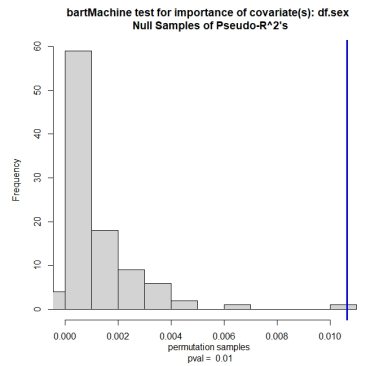
(c) sex vs. cholesterol levels

(d) sex vs. artery visibility

(e) sex vs. blood pressure

(f) sex vs. pain levels

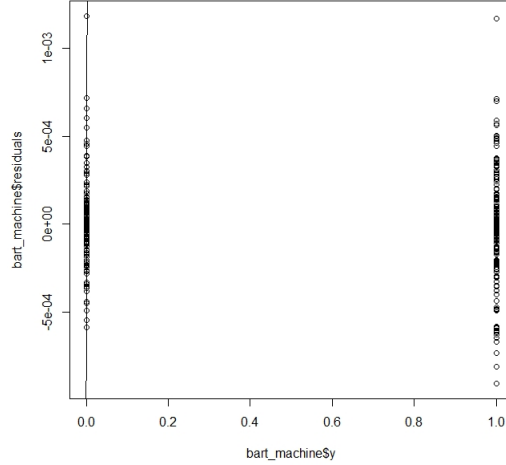Figure 4: Significance testing results with BART

Figure 5: Residual graph for the BART model

0.4.2 BART Model Performance

Figure 4 is the residual plot for the BART model. For the construction of the model, we built a train-and-test set from the binary treatment variables. This is a table of the independent variable y and its estimated response. In other words, the estimates for `target` versus the residuals for the generated BART model. This is a likely indicator of poor model-building, lack of insight into heterogeneity, or both. In this case, the Pseudo-R^2 value is 0.997, a cautiously good measure of fit. It may also be indicative of stronger bias.

0.4.3 BCF Model Performance

The results of our BCF causal analysis over the 5 most "statistically significant" heterogeneous effects relative to sex are shown in Figure 5. Here, we plot the estimate of an effect, $\hat{\tau}$, over the calculated τ. We set `nsim = 500` (indicating 500 MCMC iterations) for one round of testing and `nsim = 2000` for another, to gauge overfitting. The results indicate the presence of some heterogeneity between fields, though this fit is not necessarily perfect.

Histograms for τ relative to $\hat{\tau}$ are shown in Figure 6. The expected values overwhelmingly predict treatment versus nontreatment. The model appears to have leaned strongly towards indicating heart disease.
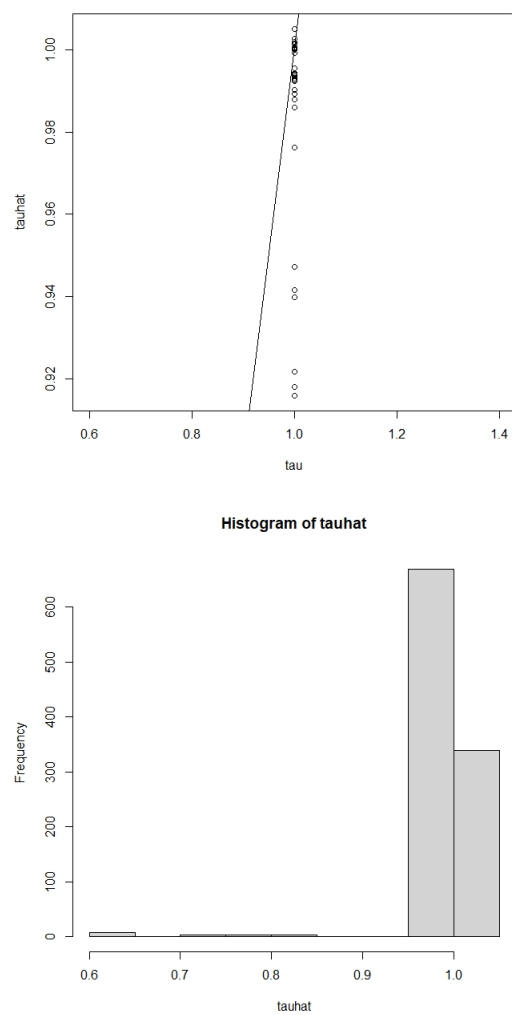
Histogram of tauhat

Figure 6: Results of BCF causal analysis performed over 5 heterogeneous effects: age, presence of heart disease, cholesterol levels, artery visibility, and blood pressure, nsim= 500 iterations
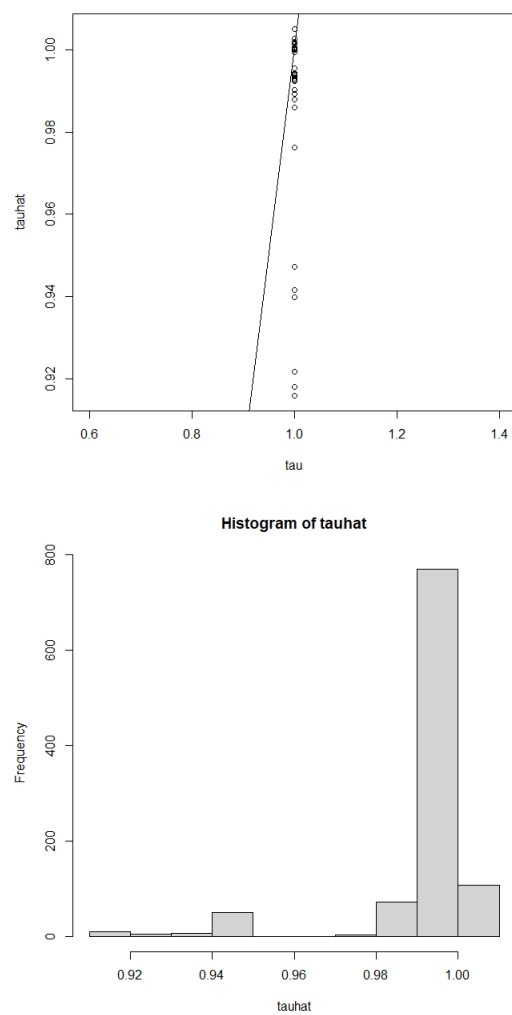
Figure 7: Results of BCF causal analysis performed over 5 heterogeneous effects: age, presence of heart disease, cholesterol levels, artery visibility, and blood pressure, nsim= 2000 iterations

0.5   Conclusions

The results from Figures 4-7 indicate that sex can be considered to be a significant predictor of heart disease. These align with the predictions of significance made by the BART model. The BART model suffered from serious overfitting, possibly due to the over-simplicity of our configuration.

Modern literature [13] [14] [15] [16] suggests that men are not only more likely to suffer from cardiovascular issues, they also begin experiencing symptoms of these issues at a younger age than women. This contrasts with the data at our disposal, which is significantly skewed to suggest that women present a heart disease at higher rates. Shortcomings like a lack of equivalent female data might present a detriment to causal estimation. However, the authors of [14] note that, while the *frequency* at which heart disease and other issues occurs in men is higher than in women, women report worse symptoms and more severe outcomes after diagnosis. The average age among women for diagnosis is generally older due to longer life expectancy [13], which is also supported by the averages computed over the dataset in Section 0.2.1, and also by the BART significance-test in Section 0.4.1.

This analysis supports the idea that women, on average, live longer before experiencing symptoms of heart disease. Further, due to the higher age range, symptoms are generally going to be more serious.

Alternatively, we could have attempted to better-fit the data to the BART model. Another potential reason for the poor residual results is in the heterogeneity of the data; BART is not necessarily always equipped to accommodate data of this nature in regression testing. Fortunately, this aligns with the reason for choosing BCF over BART in the first place.

In the future, one might perform more sophisticated analyses by choosing non-default posterior data. For instance, for simplicity, our value of $\hat{\pi}$ is just `qnorm(q)`, but more intensive analysis of the data via multiple regression testing would give a more adequate estimation. Executing a matching algorithm via the R package `matchit` would potentially give more adequate results [17]. In addition, the dataset here is relatively small, with approximately 1025 participants. The dataset used in [1] is upwards of ≈ 7000, which might help inform better training and reduce overfitting.

0.6 Bibliography

1. Hahn, P. Richard, Jared S. Murray, and Carlos M. Carvalho. "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)." Bayesian Analysis 15, no. 3 (2020): 965-1056.

2. Casella, George, and Roger L. Berger. Statistical Inference. 2nd ed. Toronto: Cengage Learning, 2002.

3. Wasserman, Larry. All of Statistics. New York: Springer, 2010.

4. "Heart Disease - Diagnosis and Treatment." Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/heart-disease/ diagnosis-treatment/drc-20353124.

5. Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). "BART: Bayesian additive regression trees." The Annals of Applied Statistics, 266–298.

6. Tan, Yaoyuan V. "Novel Applications and Extensions for Bayesian Additive Regression Trees (BART) in Prediction, Imputation, and Causal Inference." University of Michigan. Last modified , 2018. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/

7. Hahn, P. R., Puelz, D., He, J., and Carvalho, C. M. (2016). "Regularization and confounding in linear regression for treatment effect estimation." Bayesian Analysis.

8. Schwartz, Michael H., Hans Kainz, and Andrew G. Georgiadis. "Estimating Causal Treatment Effects of Femoral and Tibial Derotational Osteotomies on Foot Progression in Children with Cerebral Palsy." medrxiv.org. Last modified March 8, 2012.

   https://www.medrxiv.org/content/10.1101/2021.03.04.21252476v1.full.pdf.

9. Loyola, Juan M. "Introduction to Bayesian Additive Regression Trees." jml.github.io.

   https://jmloyola.github.io/posts/2019/06/ introduction-to-bart#

   : :text=Bayesian%20Additive%20Regression%

   20Trees%20(BART)%20is%20a%20sum%2Dof,particular%20kind%20ca

lled%20decision%20trees.

10. Kapelner, Adam, and Justin Bleich. "bartMachine: Machine Learning with Bayesian Additive Regression Trees." (2016). Journal of Statistical Software, 70(4), 1-40. doi:10.18637/jss.v070.i04

11. Prabhavathi K, Selvi KT, Poornima KN, Sarvanan A. Role of biological sex in normal cardiac function and in its disease outcome - a review. J Clin Diagn Res. 2014 Aug;8(8):BE01-4. doi: 10.7860/JCDR/2014/9635.4771. Epub 2014 Aug 20. PMID: 25302188; PMCID: PMC4190707.

12. Hahn, Richard P., Jared S. Murray, and Carlos Carvalho. "bcf: Causal Inference for a Binary Treatment and Continuous Outcome using Bayesian Causal Forests." CRAN. Last modified October 12, 2022. https://cran.r-project.org/web/packages/bcf/bcf.pdf.

13. Lam, Carolyn S., Clare Arnott, Anna L. Beale, Chanchal Chandramouli, and Denise Hilfiker-Kleiner. "Sex differences in heart failure." The European Heart Journal 40 (2019): 3859-68. https://doi.org/10.1093/eurheartj/ehz835.

14. Gao, Zujie, Zengsheng Chen, Anqiang Sun, and Xiaoyan Deng. "Gender differences in cardiovascular disease." Medicine in Novel Technology and Devices 4 (2019). https://doi.org/10.1016/j.medntd.2019.100025.

15. Mosca L, Barrett-Connor E, Wenger NK. Sex/gender differences in cardiovascular disease prevention: what a difference a decade makes. Circulation. 2011 Nov 8;124(19):2145-54. doi: 10.1161/CIRCULATIONAHA.110.968792. PMID: 22064958; PMCID: PMC3362050.

16. Fazal L, Azibani F, Vodovar N, Cohen Solal A, Delcayre C, Samuel JL. Effects of biological sex on the pathophysiology of the heart. Br J Pharmacol. 2014 Feb;171(3):555-66. doi: 10.1111/bph.12279. PMID: 23763376; PMCID: PMC3969072.

17. "R Tutorial 8: Propensity Score Matching." sejdemyr.github.io. https://sejdemyr.github.io/r-tutorials/statistics/tutorial8.htmlpropensity-score-estimation.

0.6.1   (Updated) Dataset(s) & Links

Heart Disease Dataset Link:

https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

BartPy Repository: https://github.com/JakeColtman/bartpy

bartMachine Docs: https://cran.r-project.org/web/packages/bartMachine/bartMachine.pdf

BCF Docs: https://cran.r-project.org/web/packages/bcf/bcf.pdf

GitHub Repository: https://github.com/le-kimpel/heart-disease-and-gender