# EN.625.725 Formal Methods

Lauren Kimpel

14 November 2022

**Abstract**

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.

## 1 Preliminary Analysis Methods

### 1.1 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 $Z$.

## 1.2 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 fluoroscpy, 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. These **heterogeneous effects**, with the assistance of a decent learning model, can help predict such a discrepancy.

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

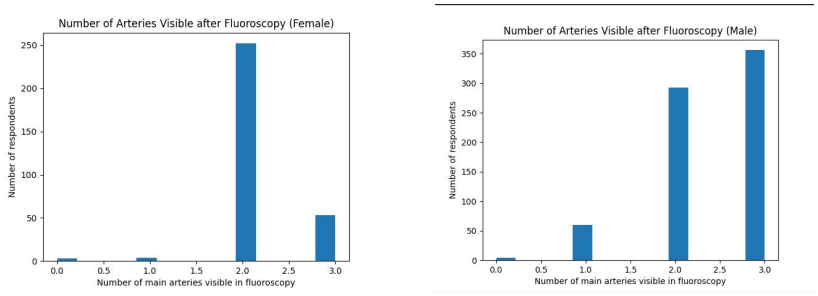
For simplicity, we will include those categories listed in Table 1 (including more if necessary.) By observation, it seems that a larger number of women than men in the dataset have outcomes which indicate heart disease, but, due to the sizeable difference in sample size between men and women, we cannot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

We will also calculate the sample variance $\hat{\sigma^2}$ and standard error $\hat{se}$ of the male versus female data for each category.

This is easy to do in python. We can initially make a surface-level comparison of the significance between each estimated parameter. This can be done using BART techniques; see Section 2.3.2 for more details on how this will be incorporated into our analysis.

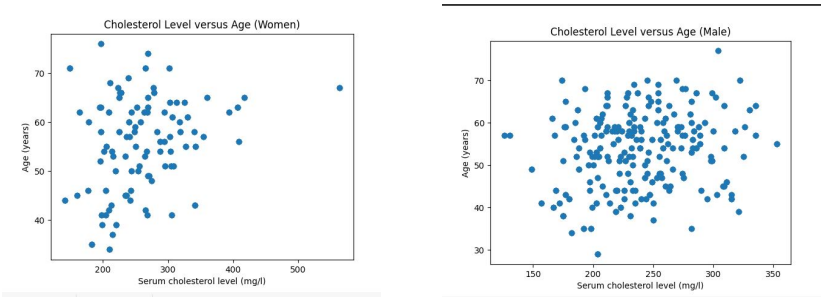
| Sample Means by Gender (m=713, f=312) | | |
|---|---|---|
| | Male | Female |
| Pain Level | 0.92 | 1.0 |
| Blood Pressure | 130.7 | 133.7 |
| Number Visible Arteries | 2.4 | 2.1 |
| Cholesterol Levels (mg/l) | 239.3 | 261.5 |
| Heart Disease True (% of Participants) | 42.1 | 72.4 |

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




(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

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

# 2 Bayesian Causal Forests

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

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

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.

## 2.1 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 `grf` and `BCF` packages. We will compare the strength of inference in both models. For simplicity, we will

employ each with default settings (the explicit prior parameterizations for `grf` 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.

## 2.2 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, $\theta_0$ 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.

### 2.2.1 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.

### 2.2.2 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 $\theta_1$ and $\theta_0$.

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

## 2.3 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 $\theta$ 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 $\beta = 2$. $\mathrm{N}^{th}$ leaves are distributed according to $\mathcal{N}(0, \sigma^2)$, with $\sigma = \frac{\sigma_0}{\sqrt{N}}$.

### 2.3.1 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 in Section 2.2. 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 $\mathcal{N}(\mu, \sigma^2)$.

### 2.3.2 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 $\beta$ 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 in R, which implements the above analysis technique.

## 2.4 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.

### 2.4.1 Fitting

Because we are adding in additional information regarding $\theta$, 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 $\theta$. Though advanced fitting options are explored in a variety of causal forest models, we will use the `grf`, `BCF`, and `bartMachine` default settings for simplicity. This is equivalent to the regularization modifications featured above.

### 2.4.2 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 1.2.1, 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 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.

# 3 Working 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.

## 3.1 (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

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

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