Conditional Relationships in the Data

Marco Morales
marco.morales@columbia.edu

Nana Yaw Essuman
nanayawce@gmail.com

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2020
Columbia University

remember this? explanation $\neq$ prediction

Explanatory Modeling	**Predictive Modeling**
f resembles $\mathcal{F}$	$\hat{f}$ links $\mathbf{X}, Y$
theory-selected $\mathbf{X}$	**association**-selected $\mathbf{X}$
may use **alternate** $\mathbf{X}$ and Y	requires **exact** $\mathbf{X}$ and Y
backward-looking	**forward**-looking
model fit validation	**predictive error** validation
$\min(Bias^2)$	$\min\{Bias^2 + Var(\hat{f}(x))\}$
$\lvert E[\hat{\beta}] - \beta\rvert \to 0$	$\min(\lvert Y_{new} - \tilde{Y}_{new}\rvert)$

A parametric perspective

what are conditional relationships in the data?

- when analyzing people and behaviors, we're not only concerned about **levels**
- we typically care about behaviors **conditional** on something else happening
  - do incumbent presidents lose elections when shark attacks increase?
- note that this is **different from "holding the rest constant"**
- can be easily computed through **multiplicative interactions**

describing a **data generating mechanism through a statistical model** (with multiplicative terms)

- we start with a simple model...

$$Y = \beta_0 + \beta_1 \mathbf{X} + \beta_2 \mathbf{Z} + \epsilon$$

- ... and add the **multiplicative interaction** term

$$Y = \beta_0 + \beta_1 \mathbf{X} + \beta_2 \mathbf{Z} + \beta_3 \mathbf{XZ} + \epsilon$$

- that now accounts for the conditional relationship between X and Z

additive and conditional models are different

- a linear **additive model** assumes a **constant effect** of X on Y

$$Y = \beta_0 + \beta_1 \mathbf{X} + \beta_2 \mathbf{Z} + \epsilon$$

- an **interactive model** assumes that the effect of X on Y **depends on the value of $\mathbf{Z}$**

$$Y = \beta_0 + \beta_1 \mathbf{X} + \beta_2 \mathbf{Z} + \beta_3 \mathbf{XZ} + \epsilon$$

when $\mathbf{Z} = 0$ (after substituting):

$$Y = \beta_0 + \beta_1 \mathbf{X} + \cancelto{0}{\beta_2 \mathbf{Z} + \beta_3 \mathbf{XZ}} + \epsilon$$

when $\mathbf{Z} = 1$ (after rearranging terms):

$$Y = [\beta_0 + \beta_2(1)] + [\beta_1 + \beta_3(1)]\mathbf{X} + \epsilon$$

additive and conditional models are different

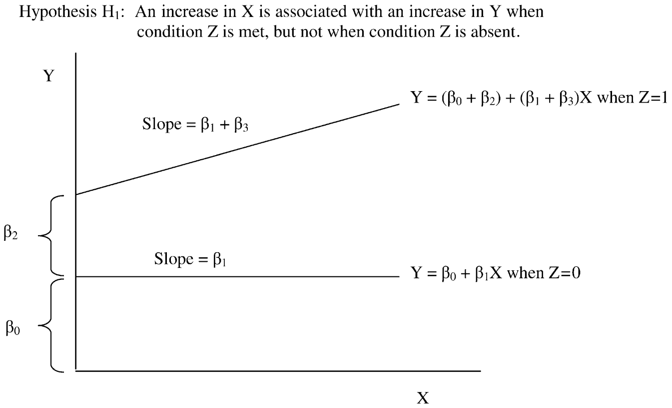


Fig. 1 A graphical illustration of an interaction model consistent with hypothesis H_1.

Figure: Brambor et al. (2006)

remember: always include **all constitutive terms**

- to estimate conditional effects **without bias**, all **lower level terms to the interaction** must be also estimated

- a **two-way interaction model** should look like:

$$Y = \beta_0 + \beta_1\mathbf{X} + \beta_2\mathbf{Z} + \beta_3\mathbf{XZ} + \epsilon$$

- a **three-way interaction model** should look like:

$$\begin{aligned} Y = \beta_0 &+ \beta_1\mathbf{X} + \beta_2\mathbf{Z} + \beta_3\mathbf{W} \\ &+ \beta_4\mathbf{XZ} + \beta_5\mathbf{XW} + \beta_6\mathbf{ZW} \\ &+ \beta_7\mathbf{XZW} + \epsilon \end{aligned}$$

remember: always include **all constitutive terms**

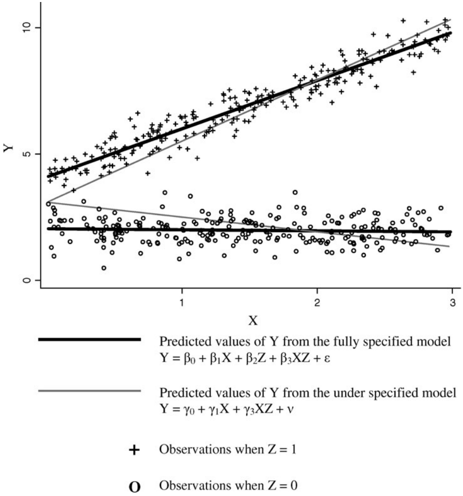


Fig. 2 An illustration of the consequences of omitting a constitutive term.

Figure: Brambor et al. (2006)

marginal effects help interpret conditional relationships

- from the (interactive) model

$$Y = \beta_0 + \beta_X \mathbf{X} + \beta_Z \mathbf{Z} + \beta_{XZ} \mathbf{XZ} + \epsilon$$

- we are interested in the **marginal effect of** X **given** Z on Y

$$\frac{\partial E[Y|X,Z]}{\partial \mathbf{X}} = \beta_X + \beta_{XZ} \mathbf{Z}$$

- it is **wrong** to assume that β_{XZ} is the **marginal effect** of X given Z on Y
    - β_X is the effect of X on Y that does not depend on Z (i.e the marginal effect when $Z = 0$)
    - β_{XZ} is the <u>part of the effect</u> of X on Y that depends on Z (when $Z \neq 0$)
- **marginal effects** of interactions are **composite quantities**

interactions also have an associated **uncertainty**

- in addition to the marginal effect

$$\frac{\partial E[Y|X,Z]}{\partial \mathbf{X}} = \beta_X + \beta_{XZ}\mathbf{Z}$$

- we need to compute its **appropriate standard error**

$$Var\left(\frac{\partial \hat{E}[Y|X,Z]}{\partial \mathbf{X}}\right) = Var[\hat{\beta}_X] + \mathbf{Z}^2 Var[\hat{\beta}_{XZ}] + 2\mathbf{Z}Cov[\hat{\beta}_X, \hat{\beta}_{XZ}]$$

An example: the Mexican war on drugs (2006-2012)

In 2016, the New York Times published an article detailing the “lethality” of Mexican armed forces in their fight against organized crime and drug cartels ongoing since 2006. Based on data released by the Mexican Government, the article concludes that

Mexico’s armed forces are exceptionally efficient killers — stacking up bodies at extraordinary rates. [...] The Mexican Army kills eight enemies for every one it wounds. [...] For the nation’s elite marine forces, the discrepancy is even more pronounced: The data they provide says they kill roughly 30 combatants for each one they injure.

Using data released by the Mexican government, we estimate an interactive model...

```
ols_interaction <-
  lm(organized_crime_dead ~ organized_crime_wounded +
       afi*long_guns_seized +
       army*long_guns_seized +
       navy*long_guns_seized +
       federal_police*long_guns_seized +
       afi*cartridge_sezied +
       army*cartridge_sezied +
       navy*cartridge_sezied +
       federal_police*cartridge_sezied +
       small_arms_seized +
       clips_seized ,
     data = AllData)
```

and use it to answer the following questions:

- **are there more expected deaths when combat is heavier?**
  - let's look at the case of events where the Navy is involved
  - we'd need to assume that more seized heavy weapons indicate heavier combat and compute

$$\beta_{navy} + \beta_{navy,long\_guns\_seized} * long\_guns\_seized$$

- **are there less expected number of deaths when no weapons are seized?**
  - let's look at the case of the Army
  - we maintain the same assumption and compute

$$\beta_{army}$$

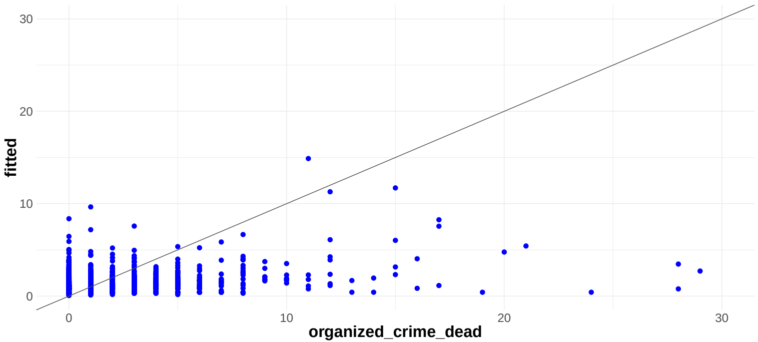
we start by looking at our estimated coefficients

```
Call:
lm(formula = organized.crime.dead ~ organized.crime.wounded +
    afi * long.guns.seized + army * long.guns.seized + navy *
    long.guns.seized + federal.police * long.guns.seized + afi *
    cartridge.sezied + army * cartridge.sezied + navy * cartridge.sezied +
    federal.police * cartridge.sezied + small.arms.seized + clips.seized,
    data = AllData)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6509 -0.7385 -0.4189  0.1933 27.2187

Residual standard error: 1.714 on 5378 degrees of freedom
Multiple R-squared:  0.1587,Adjusted R-squared:  0.156
F-statistic: 59.67 on 17 and 5378 DF,  p-value: < 2.2e-16
```

we start by looking at our estimated coefficients

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.4188645  0.0336777  12.437  < 2e-16 ***
organized.crime.wounded          0.3624050  0.0237796  15.240  < 2e-16 ***
afi                             -0.0419271  0.5040535  -0.083   0.9337    
long.guns.seized                 0.1713811  0.0172327   9.945  < 2e-16 ***
army                             0.4244453  0.0556353   7.629 2.78e-14 ***
navy                             0.2772627  0.1567621   1.769   0.0770 .  
federal.police                  -0.1113463  0.0801781  -1.389   0.1650    
cartridge.sezied                 0.0002292  0.0000968   2.368   0.0179 *  
small.arms.seized               -0.0452969  0.0186014  -2.435   0.0149 *  
clips.seized                     0.0003127  0.0003146   0.994   0.3202    
afi:long.guns.seized             0.0229013  0.0784035   0.292   0.7702    
long.guns.seized:army           -0.0459567  0.0181403  -2.533   0.0113 *  
long.guns.seized:navy            0.1761160  0.0421782   4.176 3.02e-05 ***
long.guns.seized:federal.police -0.0253811  0.0190541  -1.332   0.1829    
afi:cartridge.sezied            -0.0050516  0.0031231  -1.617   0.1058    
army:cartridge.sezied           -0.0003911  0.0000981  -3.987 6.78e-05 ***
navy:cartridge.sezied           -0.0006909  0.0001728  -3.998 6.47e-05 ***
federal.police:cartridge.sezied -0.0001518  0.0001102  -1.377   0.1685    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

as reference, this is how the interactive model fits...



back to our example

- marginal effect of 5 seized long guns on the expected number of dead on events that involve the Navy $(\beta_{navy} + \beta_{navy,long\_guns\_seized} * 5)$

1.15
[0.74, 1.56]

- marginal effect on the expected number of dead of events that involve the Army when no long guns (zero) are seized $(\beta_{army} + \beta_{army,long\_guns\_seized} * 0)$

0.42
[0.31, 0.53]

always, always, always remember...

Brambor et al. (2006)

1. Use multiplicative interaction models **whenever one's hypothesis is conditional** in nature.
2. Include **all constitutive terms** in the model specification.
3. **Do not interpret the coefficients on constitutive terms as if they are unconditional marginal effects.**
4. Do not forget to **calculate substantively meaningful marginal effects and standard errors**.

- ... or face the wrath of the stats gods!

A non-parametric perspective

other alternatives to recover conditional effects

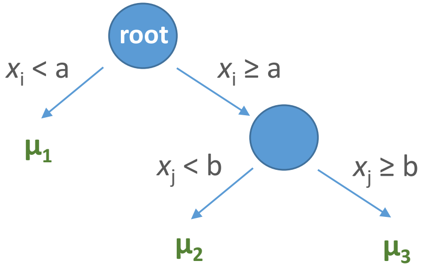
- interpreting conditional effects are a **lesser concern for prediction/classification**
  - more relevant to inferential methods that seek to describe **mechanics** of a process
- but... most **learners** can **identify interactions naturally**
  - natural candidates to capture **deep interactions**
  - **"black box"** nature **impedes direct interpretation** based on estimated parameters
  - can assess **marginal effects of $\mathbf{X}$** through **changes in predicted** Y

recap: a single tree model...

$$Y = g(x; T, M) + \epsilon$$

where

- T : a tree structure (decision rules, internal and terminal nodes)
- $M = \{\mu_1, \mu_2, ..., \mu_b\}$: set of terminal node μ's
- $g(x; T, M)$: the function that assigns a μ to x



why don't we always use single-tree models?

- single-tree models, great to account for interactions & non-linearities
  - ... but poor as predictors
- a better idea: a **sum-of-trees** model

$$
\begin{aligned}
Y = & g(x; T_1, M_1) + g(x; T_2, M_2) + \ldots + g(x; T_m, M_m) + \epsilon \\
= & \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon
\end{aligned}
$$

- each tree fits a piece of the data and "learns" from the errors of previous trees

why **Bayesian Additive Regression Trees** (BART)?

i) **sum-of-trees model**

1. fit a “weak-learning” (small) tree, and compute residuals
2. fit a new “weak-learing” tree to the residuals
3. repeat m times

ii) **regularization prior**

- maintains the depth of each tree small
- each tree contributes a small part of fit

a few characteristics of BART make it particularly useful to handle conditional relations in the data

i) naturlly identifies **interactions** and **non-linearities**
   - no subjective decisions on parametrization

ii) (virtually) **unnecessary to choose tuning parameters**
   - empirically, $m = 200$ provides good results

iii) handles a **large number of predictors**
   - naturally ignores those irrelevant for the response surface

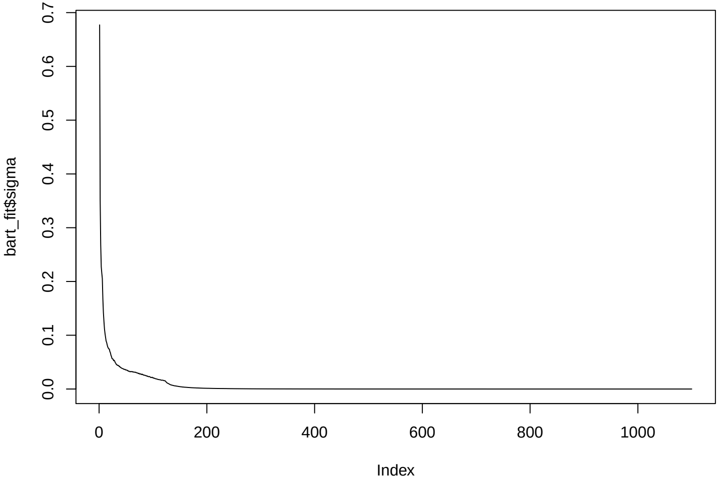
iv) straightforward to **estimate uncertainty** (from posterior distribution)

v) handles **missing data** natively

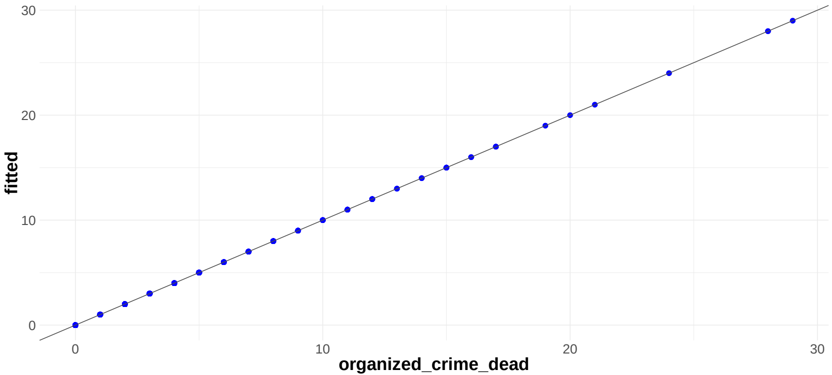
back to our example (using same predictors as before)

```
> bart_fit <- wbart(x_train, y_train)
*****Into main of wbart
*****Data:
data:n,p,np: 5396, 10, 0
y1,yn: 0.147702, -0.852298
x1,x[n*p]: 0.000000, 0.000000
*****Number of Trees: 200
*****Number of Cut Points: 16 ... 23
*****burn and ndpost: 100, 1000
*****Prior:beta,alpha,tau,nu,lambda: 2.000000,0.950000,0.512652,3.000000,0.000000
*****sigma: 0.000000
*****w (weights): 1.000000 ... 1.000000
*****Dirichlet:sparse,theta,omega,a,b,rho,augment: 0,0,1,0.5,1,10,0
*****nkeeptrain,nkeeptest,nkeeptestme,nkeeptreedraws: 1000,1000,1000,1000
*****printevery: 100
*****skiptr,skipte,skipteme,skiptreedraws: 1,1,1,1

MCMC
done 0 (out of 1100)
done 100 (out of 1100)
done 200 (out of 1100)
done 300 (out of 1100)
done 400 (out of 1100)
done 500 (out of 1100)
done 600 (out of 1100)
done 700 (out of 1100)
done 800 (out of 1100)
done 900 (out of 1100)
done 1000 (out of 1100)
time: 23s
check counts
trcnt,tecnt,temecnt,treedrawscnt: 1000,0,0,1000
```

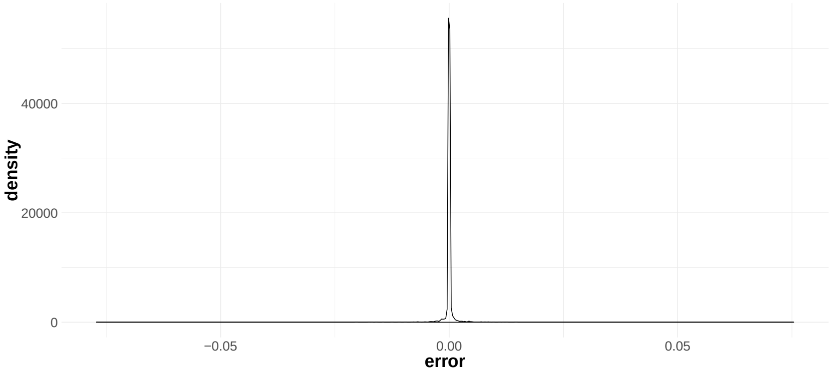
BART had a quick burn-in convergence...



BART fit the data surprisingly well...



BART produced tiny training errors



use predictive posterior to compute marginal effects

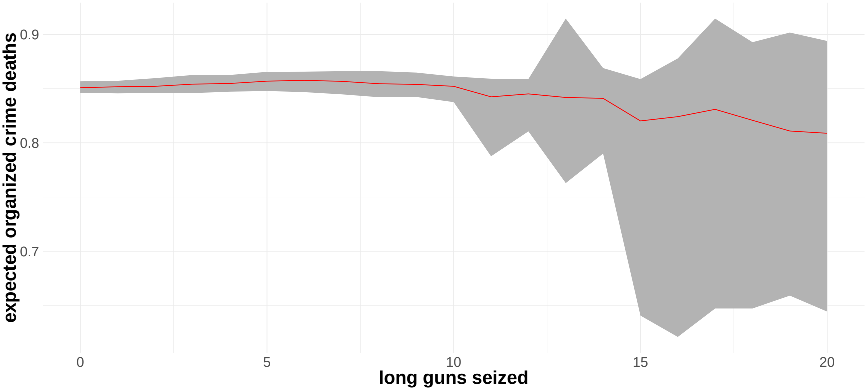
- let the training data be decomposed $x = [x_s, x_c]$, where
  - x_s: subset of covariates we intend to manipulate, and
  - x_c: complement covariates

1. set specific values for x_s in x, maintaining x_c unchanged
2. compute predicted values for new x using the predictive posterior
3. aggregate over predicted values to obtain marginal effects of x_s

$$f(x_s) = \frac{1}{N} \sum_{i=1}^{N} f(x_s, x_{ic})$$

- in our example, we set `army == 1` and `long_guns_seized` $\in [0, 20]$

use predictive posterior to compute marginal effects



Marginal effects and 95% credible intervals

Conditional Relationships in the Data

Marco Morales
marco.morales@columbia.edu

Nana Yaw Essuman
nanayawce@gmail.com

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2020
Columbia University