

# Explanation vs Prediction

Marco Morales

marco.morales@columbia.edu

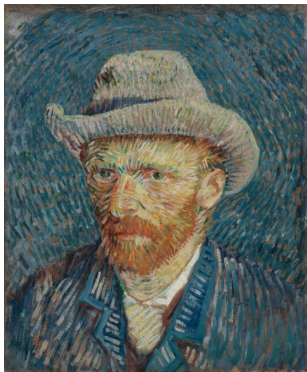
Nana Yaw Essuman

ne2388@columbia.edu

GR5069

Topics in Applied Data Science  
for Social Scientists

Spring 2025  
Columbia University



*“What matters is to grasp that which does not happen,  
in what happens”*

*- Vincent van Gogh (1882), quoted in a letter to Theo van Gogh*

# a framework to explore explanation vs prediction

Shmueli (2010)

- ▶ **theoretically...**

- ▶ let  $\mathcal{X}$  cause  $\mathcal{Y}$  through the function  $\mathcal{F}$

$$\mathcal{Y} = \mathcal{F}(\mathcal{X})$$

- ▶ **empirically...**

- ▶  $\mathbf{X}$  and  $Y$  operationalize  $\mathcal{X}$  and  $\mathcal{Y}$
  - ▶  $f$  is the model that operationalizes  $\mathcal{F}$

- ▶ **explanatory modeling** seeks an  $f$  close to  $\mathcal{F}$

$$E(Y) = f(\mathbf{X})$$

- ▶ **predictive modeling** seeks an  $\hat{f}$  that best predicts  $Y_{new}$

$$E(Y_{new}) = \hat{f}(\mathbf{X}_{new})$$

# a different — but related — perspective

Expected Prediction Error (Hastie et al. 2009)

$$EPE = \text{Var}(Y) + \text{Bias}^2 + \text{Var}(\hat{f}(x)) \quad (1)$$

where:

$EPE$  = Expected Prediction Error

$\text{Var}(Y) = E\{Y - f(x)\}^2$  : random error

$\text{Bias}^2 = \{E(\hat{f}(x)) - f(x)\}^2$  : model misspecification

$\text{Var}(\hat{f}(x)) = E\{\hat{f}(x) - E(\hat{f}(x))\}^2$  : sample estimation

## ► explanatory modeling

$$\min\{\text{Bias}^2\}$$

## ► predictive modeling

$$\min\{\text{Bias}^2 + \text{Var}(\hat{f}(x))\}$$

# in more detail: how explanation $\neq$ prediction

## Explanatory Modeling

$f$  resembles  $\mathcal{F}$

**theory**-selected  $\mathbf{X}$

may use **alternate**  $\mathbf{X}$  and  $Y$

**backward**-looking

**model fit** validation

$\min(\text{Bias}^2)$  in (??)

$$|E[\hat{\beta}] - \beta| \rightarrow 0$$

## Predictive Modeling

$\hat{f}$  links  $\mathbf{X}, Y$

**association**-selected  $\mathbf{X}$

requires **exact**  $\mathbf{X}$  and  $Y$

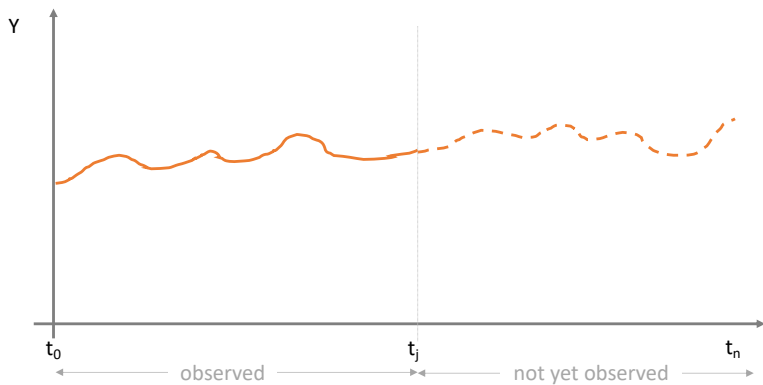
**forward**-looking

**predictive error** validation

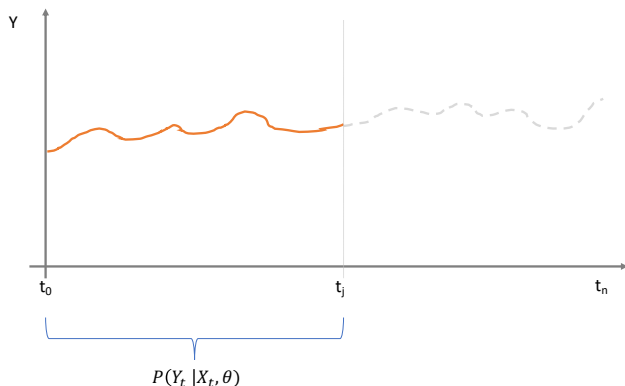
$\min\{\text{Bias}^2 + \text{Var}(\hat{f}(x))\}$  in (??)

$$\min(|Y_{\text{new}} - \tilde{Y}_{\text{new}}|)$$

## Example: think of a simple time-series

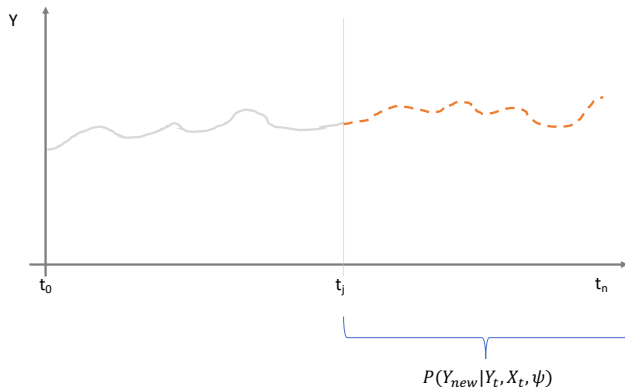


# Explanation: problem to solve is observed $Y$



- ▶ **explanatory models:** characterize  $Y$  exactly as observed — and not otherwise
- ▶ must estimate  $\theta$  w/o bias to make valid **inferences**

Prediction: problem to solve is  $Y_{new}$



- ▶ **predictive models:** project  $Y_{new}$  based on  $X_{new}$
- ▶ most likely,  $\psi \neq \theta$  where  $\psi$  exists but may be useless for inference



## to put it in perspective...

- ▶ any model will contain a combination of degrees of:
  - ▶ **explanatory power**
  - ▶ **predictive accuracy**
- ▶ two different dimensions or one with tradeoffs?
- ▶ a "good" model is **sophisticatedly simple** (Zellner 2001)

# Explanation

# what do we mean by explanation?

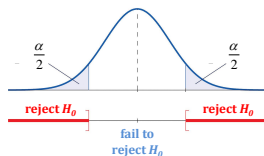
- ▶ **explanation:** provide a rationale for **why something happened the way it happened** and not in a different way
  - ▶ necessary for meaningful **inference**
  - ▶ more formally, find a **model** that describes

$$E(Y) = f(X, \theta)$$

- ▶ **problems to solve:**
  - ▶ approximate the “true” **data-generating mechanism**
    - ▶ find the  $f$  that is sufficiently close to  $\mathcal{F}$
  - ▶ recover the “true” **parameter values** that govern the observed data
    - ▶ find the appropriate  $\theta$  in  $P(Y|X, \theta)$

# focus for inference: successful recovery of $\theta$

- ▶ inference requires having the “**correct**” **estimated parameters**  $\hat{\theta}$ , so we spend time
  - i) carefully **describing null hypotheses** for estimated parameters:  $H_0 : \theta = 0$
  - ii) **characterizing distributions under the null**:  $\theta \stackrel{H_0}{\sim} t_v(0, 1)$
  - iii) evaluating **how unlikely** is the estimated value under the null:  $Pr(\hat{\theta} > k)$



- iv) using theory to ensure that estimated parameters are **unbiased**
- v) using statistics to ensure that **appropriate variances** were estimated

# an example: price elasticity of demand

- ▶ based on economic theory



- ▶ we hypothesize a relation between price ( $p$ ) and quantity sold ( $q$ ) of a good — **price elasticity**:

$$\eta_D = \frac{\% \Delta q^D}{\% \Delta p}$$

- ▶ we fit a model and estimate  $\hat{\beta}$  to recover  $\eta_D$

$$\log(q) = \alpha + \beta * \log(p) + \epsilon$$

- ▶ from the estimated parameter  $\hat{\beta}$ , we infer that a 1% increase in  $p$  decreases  $q$  by 0.4%

# do not forget: experiments and causal inference

- ▶ recovering parameters for inference is hard — too many things could go wrong
  - ▶ **observational data** is messy; must rely on theory as guidance
- ▶ **causal inference** is even harder: must **compare potential outcomes** under treatment and without treatment
  - ▶ **experiments** are golden standard for **causal inference**: treatment assignment is **unconfounded**
  - ▶ **observational data** for causal inference requires balance in pre-treatment covariates to compare appropriate groups
    - ▶ observational data does not always have “untreated” observations

# Prediction

# what do we mean by prediction?

- ▶ etimologically:
  - ▶ ***predict***: prae- before + dicere to say
  - ▶ ***forecast***: fore- before + casten to prepare
  - ▶ ***prognosticate***: pro- before + gnoscerere to know
- ▶ generically, the use of a **model** that leverages **observed information** to project **new information**

$$\tilde{Y}_{new} = \hat{f}(Y_{obs}, X_{obs})$$

- ▶ **problem to solve**: find an “appropriate” model  $\hat{f}$  that can produce  $\tilde{Y}_{new}$  with small errors ( $\min\{|Y_{new} - \tilde{Y}_{new}|\}$ )



# some empirical considerations for prediction

- ▶ **Predictability** depends on (Hyndman et al. 2013):
  - ▶ how well we know factors that influence the predictions
  - ▶ how much data (and of what quality!)
  - ▶ recursive influence of predictions (especially forecasts)
- ▶ **Key question: what to predict?**
  - ▶ every item?
  - ▶ at what level of aggregation?
  - ▶ at what frequency?
- ▶ **Objective:** find a model with **consistently “small” predictive errors**
  - ▶ cope with risk of **overfitting** the model

# what do we mean by “overfitting”?

- ▶ **overfitting:** capturing **patterns in the training data** that do not extend to new observations
- ▶ an overfit model may generate **systematically large predictive errors**
  - ▶ predictions are **not generalizable** to new data
- ▶ **challenge:** find the **set of predictors** that carry the appropriate “signal” to projections of the future
  - ▶ enough information to capture **meaningful patterns**
  - ▶ ...not so much as to also capture patterns that are **irrelevant for the future**
  - ▶ the **bias-variance tradeoff**

# overfitting: the bias-variance tradeoff perspective

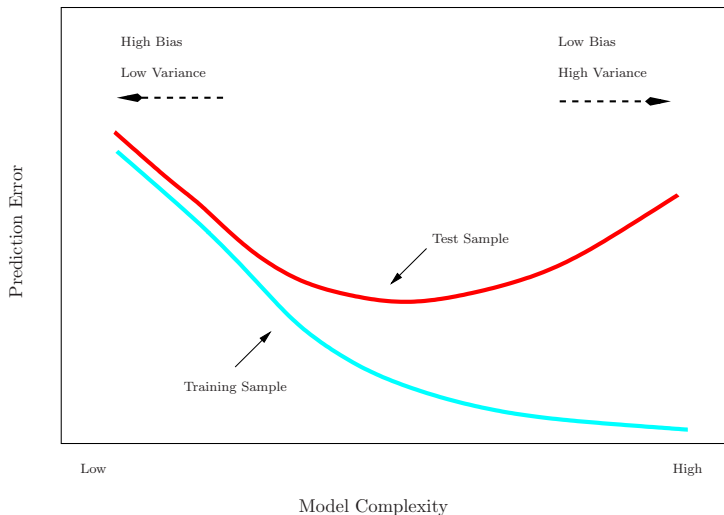
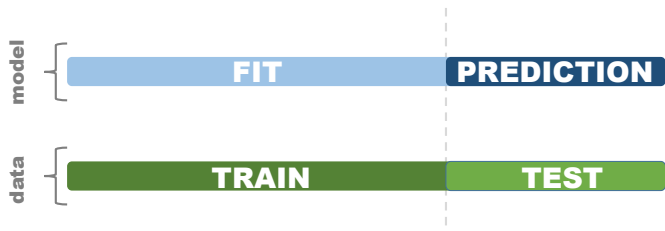


Figure: James et al (2013)

# validation to minimize overfitting

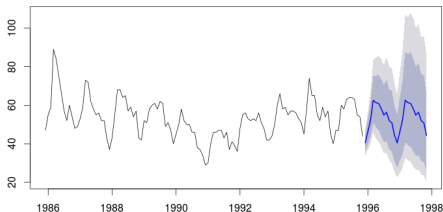
- i) fit model on a **training set**
- ii) measure error on a **test set**



- ▶ usually, error on **test set** > error on **training set**
  - ▶ **caveat:** training and test sets should come **from the same population**
- ▶ **validation** can take many flavors (e.g. **k-fold validation**, **leave p-out cross-validation...**)

# do not forget: predictions carry uncertainty!

- ▶ the future is unknown, therefore **predictions have uncertainty**
  - ▶ many predictive models only produce **point estimates** of predictions, and ignore **prediction intervals**
  - ▶ may generate **erroneous impression** that **predictions have no uncertainty**
- ▶ when possible, estimate the **range of values** where predictions may lie **with a given probability**



# some (empirically validated) rules of thumb

## 1. **keep it simple:**

- ▶ start parsimonious and add complexity (*iff* called for)
- ▶ increased complexity typically reduces accuracy

## 2. **rely on domain expertise to select inputs**

- ▶ statistical significance a faulty guide for inclusion
- ▶ domain expertise should drive variables to include

## 3. **include more (useful) information**

- ▶ high correlation in predictors not an issue

## 4. **fit $\neq$ accuracy**

- ▶ well-fitting models may impose unwarranted “structure” and “certainty” to the forecast

## 5. **update models constantly**

- ▶ update parameters as new information arrives

# **ADDENDUM:**

## **Conditional Relations in the Data**

# 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

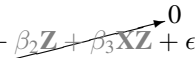
- ▶ a linear **additive model** assumes a **constant effect** of  $X$  on  $Y$

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

- ▶ an **interactive model** assumes that the effect of  $X$  on  $Y$  **depends on the value of  $Z$**

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

when  $Z = 0$  (after substituting):

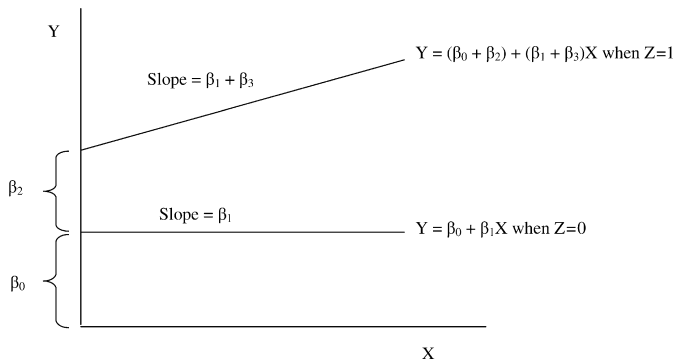
$$Y = \beta_0 + \beta_1 \mathbf{X} + \beta_2 \mathbf{Z} + \beta_3 \mathbf{XZ} + \epsilon$$


when  $Z = 1$  (after rearranging terms):

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

# additive and conditional models are different

Hypothesis  $H_1$ : An increase in  $X$  is associated with an increase in  $Y$  when condition  $Z$  is met, but not when condition  $Z$  is absent.



**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**

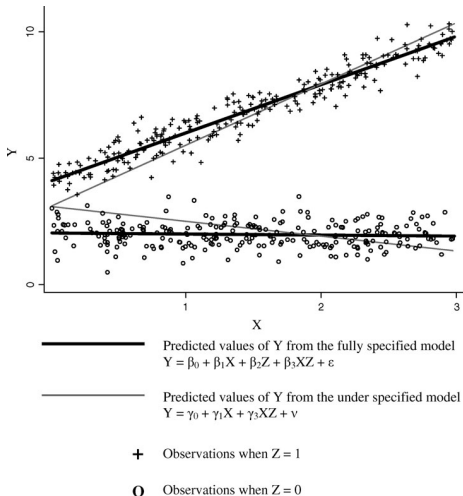
- ▶ to estimate conditional effects **without bias**, all **lower level terms to the interaction** must be also estimated
- ▶ a **two-way interaction model** should look like:

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

- ▶ 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  $\beta_{XZ}$  is the **marginal effect** of X given Z on Y
  - ▶  $\beta_X$  is the effect of X on Y that does not depend on Z (i.e the marginal effect when  $Z = 0$ )
  - ▶  $\beta_{XZ}$  is the part of the effect 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_seized +  
      army*cartridge_seized +  
      navy*cartridge_seized +  
      federal_police*cartridge_seized +  
      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.seized + army * cartridge.seized + navy * cartridge.seized +
    federal.police * cartridge.seized + 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

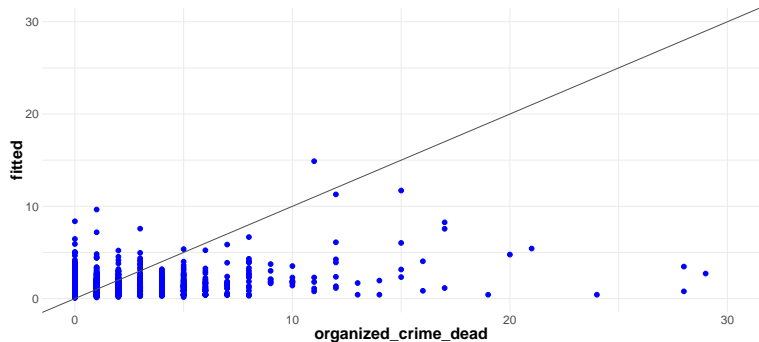
Coefficients:

	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.seized	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.seized	-0.0050516	0.0031231	-1.617	0.1058	
army:cartridge.seized	-0.0003911	0.0000981	-3.987	6.78e-05	***
navy:cartridge.seized	-0.0006909	0.0001728	-3.998	6.47e-05	***
federal.police:cartridge.seized	-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

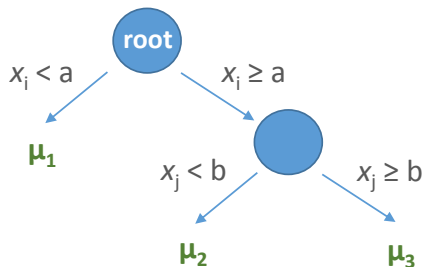
- ▶ interpreting conditional effects are a **lesser concern for prediction/classification**
  - ▶ more relevant to inferential methods that seek to describe **mechanics** of a process
- ▶ but... most **learners** can **identify interactions naturally**
  - ▶ natural candidates to capture **deep interactions**
  - ▶ “**black box**” nature **impedes direct interpretation** based on estimated parameters
  - ▶ can assess **marginal effects of  $X$**  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)
- ▶  $M = \{\mu_1, \mu_2, \dots, \mu_b\}$  : set of terminal node  $\mu$ 's
- ▶  $g(x; T, M)$  : the function that assigns a  $\mu$  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

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

- ▶ 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-learning” 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

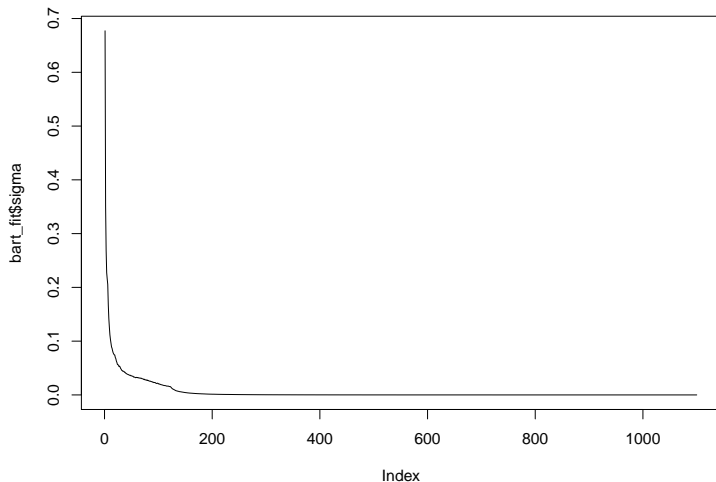
- i) naturally identifies **interactions** and **non-linearities**
  - ▶ no subjective decisions on parametrization
- ii) (virtually) **unnecessary to choose tuning parameters**
  - ▶ empirically,  $m = 200$  provides good results
- iii) handles a **large number of predictors**
  - ▶ naturally ignores 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
yl,yn: 0.147702, -0.852298
xl,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,nkeepstest,nkeepstestme,nkeepstreedraws: 1000,1000,1000,1000
*****printevery: 100
*****skiptr,skipte,skipteme,skiptreedraws: 1,1,1,1
```

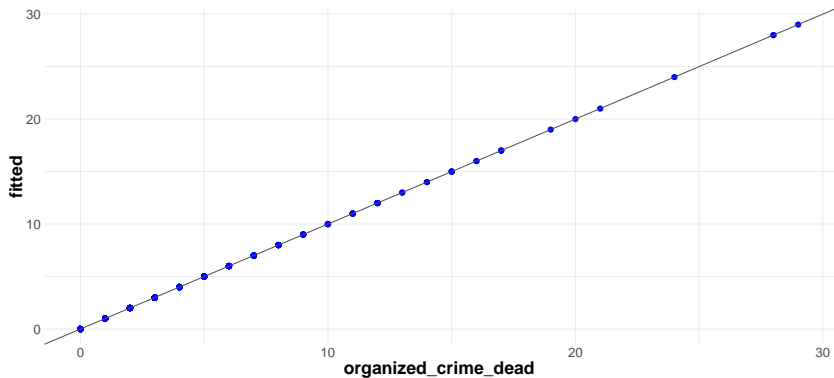
```
MCMC
done 0 (out of 1100)
done 100 (out of 1100)
done 200 (out of 1100)
done 300 (out of 1100)
done 400 (out of 1100)
done 500 (out of 1100)
done 600 (out of 1100)
done 700 (out of 1100)
done 800 (out of 1100)
done 900 (out of 1100)
done 1000 (out of 1100)
time: 23s
check counts
trcnt,tecnt,temecnt,treedrawscnt: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

# BART had a quick burn-in convergence...

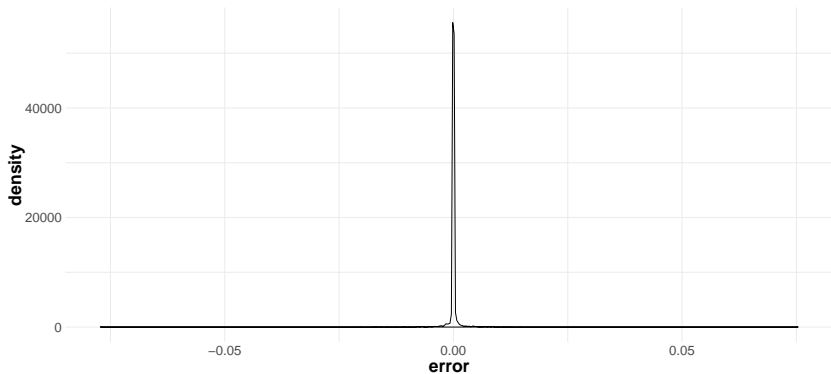




BART fit the data surprisingly well...



# BART produced tiny training errors



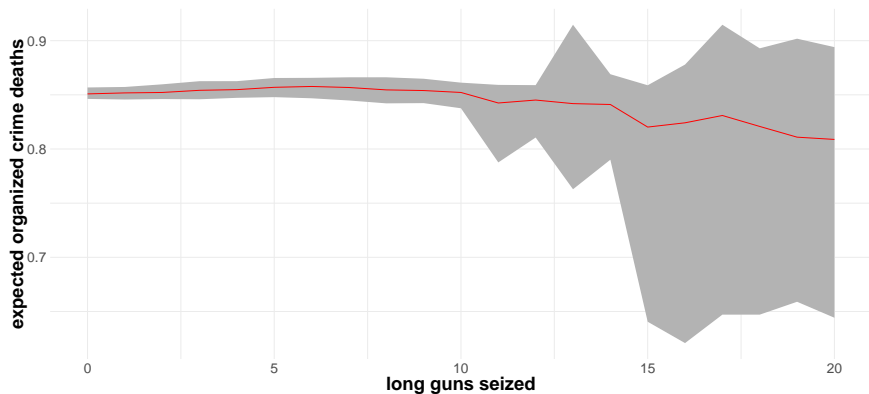
# use predictive posterior to compute marginal effects

- ▶ 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 ∈ [0, 20]`

# use predictive posterior to compute marginal effects



Marginal effects and 95% credible intervals

# Explanation vs Prediction

Marco Morales

marco.morales@columbia.edu

Nana Yaw Essuman

ne2388@columbia.edu

GR5069

Topics in Applied Data Science  
for Social Scientists

Spring 2025  
Columbia University