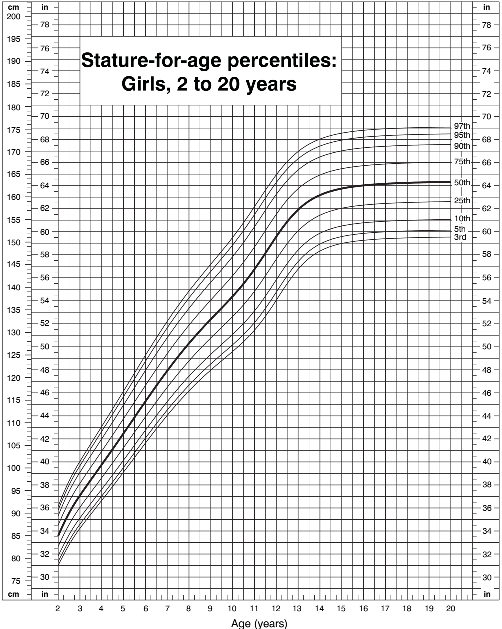
Machine learning regression and marginal effects inference

Rodney Sparapani
Associate Professor of Biostatistics
Medical College of Wisconsin

Marquette University Colloquium

CDC Growth Charts: United States



Motivating Example: Growth Charts

- US Centers for Disease Control and Prevention (CDC) and the World Health Organization have developed growth charts for childhood development: height by age, weight by age, body mass index by age and weight by height
- Here we will focus on height, y_t, by age in months, $t = 24, \ldots, 215$ (2 to 17 years old)
- CDC uses the LMS method via natural cubic splines (Cole and Green 1992 *Statistics in Medicine*)
- Three parameters estimated by penalized maximum likelihood the Box-Cox power transformation, L_t; the mean, M_t; and the coefficient of variation, S_t

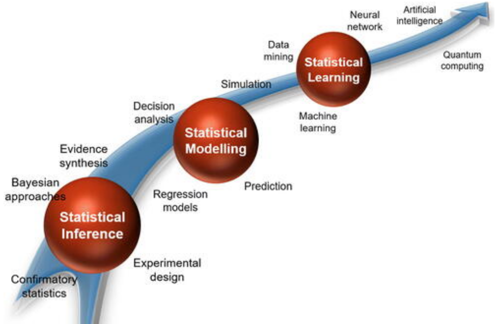
$$z_t = \left\{ \begin{array}{ll} \frac{-1+(y_t/M_t)^{L_t}}{L_t S_t} & L_t \neq 0 \\ \frac{\log(y_t/M_t)}{S_t} & L_t = 0 \end{array} \right\} \sim \mathrm{N}(0,\ 1)$$

- But, this only uses part of the data: just males or just females
- What if we wanted to use all of the data?
- Or include more information like weight and race/ethnicity?

What is Artificial Intelligence and Statistical Learning?

Artificial intelligence (AI) is a computer system's ability to perform tasks that normally require human intelligence such as driving a car

- 1941 (circa): "Machine Intelligence" coined by Alan Turing
- 1950: Turing's *Imitation Game* (alike today's *Turing Test*)
- 1956: "Artificial Intelligence" coined at Dartmouth Workshop
- 1950 to 2010: AI 1.0, basic research with limited capabilities
- 2011 to 2017: AI 2.0, deep learning
- 2018 to today: AI 3.0, foundation/large-language models
- Howell, Corrado & DeSalvo 2024 *JAMA*



What is Machine Learning (or Statistical Learning)?

- *Machine learning*, or statistical learning, is a field within AI to develop methods that learn statistical relationships from *training data* without being explicitly programmed to do so (paraphrasing computer scientist Arthur Samuel 1959)
- For example, you could physically model childhood growth chart data based on principles of human auxology or you could nonparametrically learn the growth curves from training data
- Back in Samuel's day, linear/logistic regression were considered *machine learning regression (MLR)* for lack of alternatives; however, they do NOT meet the definition due to restrictive linearity and precarious parametric assumptions
- Linear/logistic regression are proto-MLR rather than MLR
- Today, by the term "MLR", I mean the widely flexible sense of without being explicitly programmed to do so

What are black-box models?

- The term *black-box*, coined in 1945, for the development of an experimental analysis with electronic circuits that had been in practice about 20 years at that time (Belevitch 1962)
- Simply ignore the circuit details as-if hidden inside a **black-box** instead, characterize the response output from its stimulus input via experimentation, trial and error, etc.
- MLR's are typically black-boxes and that is a down-side
  a direct interpretation of the model itself is not evident
  due to complexity, so don't even bother trying (in stark contrast to the trivial linear/logistic regression coefficients)
- In modern terms, a black-box model defies understanding via inspection of the covariates and their associated parameters
- Rather, an intuitive interpretation is devised by other means such as an orchestrated sequence of covariate setting predictions
- Therefore, the rising interest in marginal *(explainable)* effects
- Marginal effects are applicable to MLR in general, but here our focus is on Bayesian Additive Regression Trees (BART)

What is Machine Learning Regression (MLR)?

- MLR is extensible, but for the moment consider the general regression case of a continuous outcome with Normal errors

$$y_i = \mu_0 + f(x_i) + \epsilon_i \qquad \text{where } \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}\left(0,\ \sigma^2\right)$$

- f is an unspecified function whose form is to be *learned* from the training data and x_i is a vector of covariates for $i = 1, \ldots, N$
- An important modern MLR extension that we will only touch on

$$y_i = \mu_0 + f(x_i) + s(x_i)\epsilon_i \qquad \text{where } \epsilon_i \overset{\text{iid}}{\sim} F_\epsilon$$

- f alone (or f and s) will be *learned*, but how?
- Following Samuel's principle via Bayesian nonparametric models without resorting to precarious restrictive assumptions we don't want to assume linearity nor pre-specify interactions

What is Machine Learning Regression (MLR)?

- *Ensemble learning* discovered in 1997
  Krogh & Solich 1997 *Physical Review E*
- An ensemble of *machines* (in our case binary trees) are fit simultaneously that form the basis of an aggregate prediction with superior performance to any single machine's fit
- Ensembles are the best currently-known machine learning method with respect to out-of-sample predictive performance for so-called *tabular data* where all of the covariates are of different types, i.e., age, sex, height, weight, etc.
- N.B. *Deep learning* is inferior to ensembles for tabular data
  for optimal artificial neural net performance, the inputs need to be all the same type, i.e., all pixels, words or audio waves, etc.
  Lundberg and Erion et al. 2020 *Nature Machine Intelligence*
  Shwartz-Ziv and Armon 2022. *Information Fusion*

Why are Ensemble Learning predictions optimal?

- There is a trade-off between the bias and variance
- mean squared error = bias^2 + variance
- Consider the spectrum of trade-offs

  Linear regression is on the high bias/low variance end

  Single-tree regression is on the low bias/high variance end
- While ensemble are in between: medium bias/medium variance
- BART is in the class of ensembles that both theoretically, and in practice, have optimal out-of-sample predictive performance

Baldi & Brunak 2001 "Bioinformatics: machine learning approach"
Kuhn & Johnson 2013 "Applied Predictive Modeling"

Selected BART references with URLs

Inception	Chipman, George & McCulloch 2010 *AOAS*
BART R package	Sparapani, Spanbauer & McCulloch 2021 *JSS*
Heteroskedastic	Chipman, George et al. 2021 *Bayesian Analysis*
Monotonicity & Outlier Detection	Pratola, Chipman et al. 2020 *JCGS* Sparapani, Teng et al. 2022 *JPGN*
Variable Selection (Big P)	Linero 2018 *JASA* Liu, Rockova 2023 *JASA*
Big Data (Big N)	Pratola, Chipman et al. 2014 *JCGS* Entezari, Craiu et al. 2017 *Canadian J of Stat*
Skew/Multivariate	Um, Linero et al. 2023 *Statistics in Medicine*
Nonparametric Theory	Rockova & Saha 2019 *PMLR* Rockova & van der Pas 2020 *AOS*
Survival Analysis	Sparapani, Logan et al. 2016 *Statistics in Medicine* Sparapani, Rein et al. 2020 *Biostatistics* Sparapani, Logan et al. 2020 *SMMR* Linero, Basak et al. 2021 *Bayesian Analysis* Sparapani, Logan et al. 2023 *Biometrics*

Single-tree regression model

Chipman, George & McCulloch 1998 *JASA*

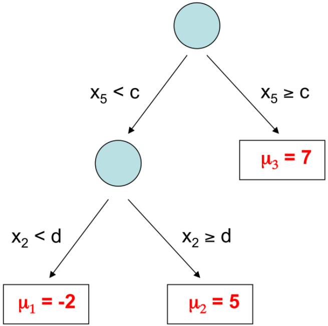
y_i is a continuous outcome where i indexes subjects $i = 1, \ldots, N$

x_i is a vector of covariates

$\mathcal{T}$ denotes the tree structure and branch decision rules

$\mathcal{M} \equiv \{\mu_1, \mu_2, \ldots, \mu_L\}$ denotes the leaf values

$g(x_i; \mathcal{T}, \mathcal{M})$ is a regression tree function



$$y_i = \mu_0 + g(x_i; \mathcal{T}, \mathcal{M}) + \epsilon_i \text{ where } \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$$

Bayesian Additive Regression Trees (BART)

Chipman, George & McCulloch 2010 *Annals of Applied Stat*

$$y_i = \mu_0 + f(x_i) + \epsilon_i \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}\left(0,\ w_i^2\sigma^2\right)$$

$$f \overset{\text{prior}}{\sim} \mathbf{BART}\ (a, b, H, \kappa, \mu_0, \tau)$$

$$f(x_i) \equiv \sum_{h=1}^{H} g(x_i; \mathcal{T}_h, \mathcal{M}_h) \qquad H \in \{50, 200, 500\}$$

$$\mu_{hl}|\mathcal{T}_h \overset{\text{prior}}{\sim} \mathrm{N}\left(0,\ \frac{\tau^2}{4H\kappa^2}\right) \text{ leaves of } \mathcal{T}_h$$

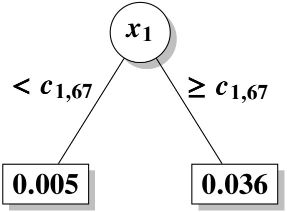
$$\in \mathcal{M}_h$$

$$\sigma^2 \overset{\text{prior}}{\sim} \lambda\nu\chi^{-2}(\nu)$$

The **BART** R package and binary trees

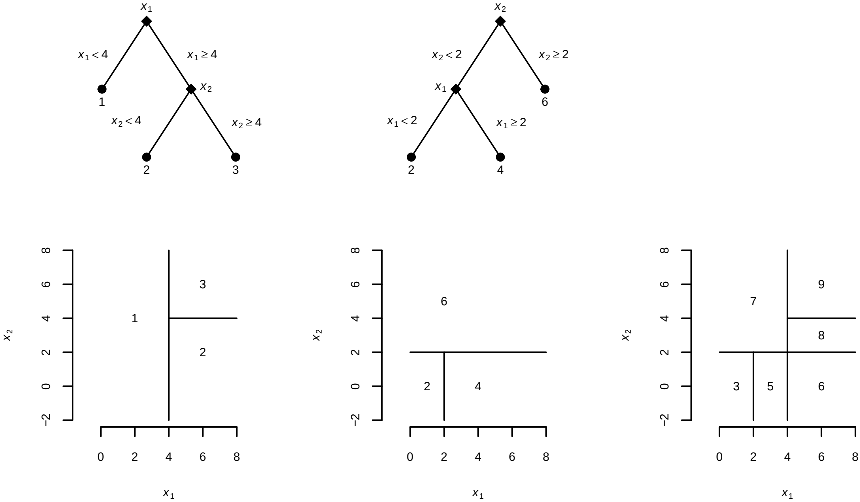
Sparapani, Spanbauer & McCulloch 2021
Journal of Statistical Software

```
R> write(post$treedraws$trees, "trees.txt")
R> tc <- textConnection(post$treedraws$tree)
R> trees <- read.table(file=tc, fill=TRUE, row.names=NULL,
+    col.names=c("node", "var", "cut", "leaf"))
R> close(tc)
R> head(trees)
  node var cut leaf
1 1000 200   1   NA
2    3  NA  NA   NA
3    1   0  66 -0.0010
4    2   0   0  0.0048
5    3   0   0  0.0357
6    3  NA  NA   NA
```



Bayesian Additive Regression Trees (BART)

Logan, Sparapani, McCulloch & Laud 2020 *SMMR*



The BART short-hand implies the following priors

Priors	
Covariate choice	$\mathbf{U}(\{1,\ldots,P\})$ or
	$\mathbf{D}\,(\theta/P,\ldots,\theta/P)$ Linero 2018 *JASA*
Branch decision point	$\mathbf{U}(\{1,\ldots,C\})$
Branching penalty	$\mathbf{P}[\mathbf{Branch}\vert\text{tier}] = a\,(1+\text{tier})^{-b}$
Default prior settings	
$a = 0.95, b = 2$	

Number of leaves	1	2	3	4+
Prior probability	0.05	0.55	0.27	0.13

BART and Bayesian nonparametric theory

- frequentist theoretical justification for BART's performance: asymptotically consistent with a near optimal learning rate
- the BART posterior distribution concentrates around the truth at a near optimal minimax rate
- the default BART Branching penalty is near optimal: $\mathbf{P[Branch|tier]} = a(1 + \text{tier})^{-b}$
- the optimal BART Branching penalty is now known to be: $\mathbf{P[Branch|tier]} = \gamma^{\text{tier}}$ **where** $0 < \gamma < 0.5$

Number of leaves	1	2	3	4+
Prior probability	0.00	$(1-\gamma)^2$	$2\gamma(1-\gamma)(1-\gamma^2)^2$	...
$\gamma = 0.25$	0.00	0.56	0.33	0.11
$a = 0.95, b = 2$	0.05	0.55	0.27	0.13

Rockova & van der Pas 2019 *Annals of Statistics*
Rockova & Saha 2019 *Proceedings of Machine Learning Research*

Marginal Effects and
Machine Learning Regression (MLR)

- Suppose we have an MLR, $f(x)$, that is likely a complex function of the covariates with nonlinearities and interactions
- And we divide the covariates into those of interest, S, and the complement, C, not of interest: $f(x) \equiv f(x_S, x_C)$
- Typically, S is of low-dimension since we intend to peak inside the black-box by visualization: usually 1 to 3 dimensions
- Let $f_S(x_S)$ denote the marginal effect of x_S

$$\mathrm{E}[y|x_S] \equiv \mu_0 + f_S(x_S)$$

$$f_S(x_S) \equiv \mathrm{E}_{x_C}[f(x_S, x_C)|x_S]$$

$$= \int \cdots \int f(x_S, x_C)[x_C|x_S]\,\mathrm{d}x_C$$

where $[x_C|x_S]$ is the distribution of $x_C|x_S$

$$= \int \cdots \int f(x_S, x_C)[x_C]\,\mathrm{d}x_C \qquad \text{assuming } x_S \perp x_C$$

Friedman's partial dependence function (FPD) and Marginal Effects Assuming Independent Covariates

$$\begin{aligned} \mathbf{E}[y|x_S] &\equiv \mu_0 + f_S(x_S) \\ f_S(x_S) &\equiv \mathbf{E}_{x_C}[f(x_S, x_C)|x_S] \\ &= N^{-1} \sum_i f(x_S, x_{iC}) \end{aligned}$$

the partial dependence function

where x_{iC} are the training values

$$\begin{aligned} f_{S_m}(x_S) &= N^{-1} \sum_i f_m(x_S, x_{iC}) \\ \hat{f}_S(x_S) &= M^{-1} \sum_m f_{S_m}(x_S) \end{aligned}$$

Friedman 2001 *Annals of Statistics*

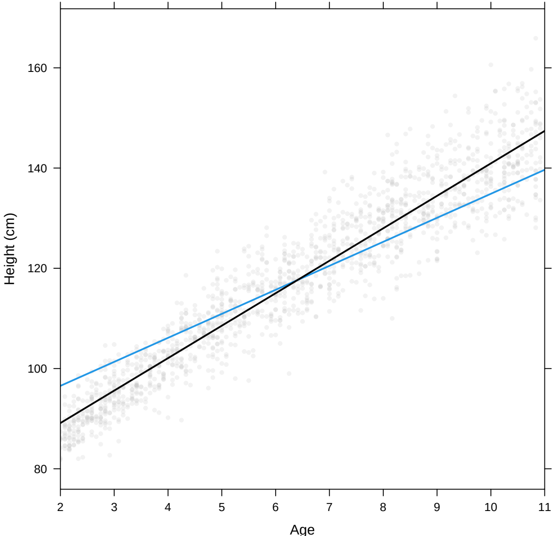
Friedman's partial dependence function (FPD) and Marginal Effects Assuming Independent Covariates

Linear regression example

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i \\
f(x_{1i}, x_{2i}) &= \beta_1 x_{1i} + \beta_2 x_{2i} \\
x_S &= x_1 \\
x_C &= x_2 \\
f_S(x_1) &= \mathbf{E}_{x_2}\left[f(x_1, x_{2i}) | x_1\right] \\
&= \mathbf{E}_{x_2}\left[\beta_1 x_1 + \beta_2 x_{2i} | x_1\right] \\
&= \beta_1 x_1 + \beta_2 \mathbf{E}_{x_2}\left[x_{2i}\right] \\
&= N^{-1} \sum_i f(x_1, x_{2i}) \\
&= N^{-1} \sum_i (\beta_1 x_1 + \beta_2 x_{2i}) \\
&= \beta_1 x_1 + \beta_2 \bar{x}_2
\end{aligned}
$$

Friedman's partial dependence function (FPD) and Marginal Effects Assuming Independent Covariates

Linear regression example



Probit BART for dichotomous outcomes

$$y_i | p_i \overset{\text{ind}}{\sim} \mathrm{B}(p_i)$$

$$p_i | f = \Phi(\mu_0 + f(x_i)) \text{ where } f \overset{\text{prior}}{\sim} \text{BART} \text{ and } \mu_0 = \Phi^{-1}(\bar{y})$$

$$z_i | y_i, f \sim \mathrm{N}(\mu_0 + f(x_i),\ 1) \begin{cases} \mathrm{I}(-\infty, 0) & \text{if } y_i = 0 \\ \mathrm{I}(0, \infty) & \text{if } y_i = 1 \end{cases}$$

$$f | z_i, y_i \overset{d}{=} f | z_i$$

Continuous BART with unit variance, $\sigma^2 = 1$ where z_i are the data
Albert & Chib 1993 *JASA*

Friedman's partial dependence function (FPD) and Marginal Effects Assuming Independent Covariates Probit BART

$$
\begin{aligned}
p(x) &= p(x_S, x_C) \\
&= \Phi(\mu_0 + f(x_S, x_C)) \\
p_S(x_S) &= \mathrm{E}_{x_C}\left[p(x_S, x_C) | x_S\right] \\
&\approx N^{-1} \sum_i p(x_S, x_{iC}) \\
&\equiv N^{-1} \sum_i \Phi(\mu_0 + f(x_S, x_{iC})) \\
p_{S_m}(x_S) &\equiv N^{-1} \sum_i p_m(x_S, x_{iC}) \\
\hat{p}_S(x_S) &\equiv M^{-1} \sum_m p_{S_m}(x_S)
\end{aligned}
$$

Extending FPD to Dependent Covariates
by the Empirical Imputation Marginal

- Consider our growth chart for height example
- To do this the right way, first we model the strong relationship between age, sex and weight among children
  $\mathrm{E}[w|t,u] = \tilde{w} = \tilde{\mu} + \tilde{f}(t,u)$
- We can summarize the relationship with a BART model
  $w_i = \tilde{\mu} + \tilde{f}(t_i,u_i) + \tilde{\epsilon}_i$ **where** $\tilde{f} \overset{\text{prior}}{\sim}$ **BART**
- For marginal effects more applicable to dependent variables

$$\begin{aligned} f_{t,u}(t,u) &= \mathrm{E}_v[f(t,u,v,\tilde{w})|t,u,\tilde{w}] \\ &= K^{-1}N^{-1}\sum_k \sum_i f(t,u,v_i,\tilde{f}_k(t,u)) \end{aligned}$$

where $\tilde{f}_k(t,u)$ are draws from the posterior

- However, this is just performing FPD K times
  so K times more computationally demanding!

Extending FPD to Dependent Covariates

by the Direct Imputation Marginal

- Consider our growth chart for height example
- We proceed as before by modeling the strong relationship between age, sex and weight among children
  $\mathbf{E}\,[w|t,u] = \tilde{w} = \tilde{\mu} + \tilde{f}(t,u)$
  $w_i = \tilde{\mu} + \tilde{f}(t_i,u_i) + \tilde{\epsilon}_i$ where $\tilde{f} \overset{\text{prior}}{\sim}$ BART
- For marginal effects more applicable to dependent variables

$$
\begin{aligned}
f_{t,u}(t,u) &= \mathrm{E}_v\,[f(t,u,v,\tilde{w})|t,u,\tilde{w} = \mathrm{E}[w|t,u]] \\
&= \mathrm{E}_v\left[f(t,u,v,\tilde{f}(t,u))|t,u\right] \\
&= N^{-1}\sum_i f(t,u,v_i,\tilde{f}_*(t,u)) \\
&\text{where } \tilde{f}_*(t,u) = M^{-1}\sum_m \tilde{f}_m(t,u)
\end{aligned}
$$

- Now this is just performing FPD one time so a more friendly computation!

Extending FPD to Dependent Covariates
by the Nearest Neighbor Marginal

- Again consider our growth chart for height example
- t for age, u for sex, v for race/ethnicity and w for weight
- For age, t, we have a carefully chosen grid of values $-\infty = \tilde{t}_0 < \tilde{t}_1 < \tilde{t}_2 < \cdots < \tilde{t}_J < \tilde{t}_{J+1} = \infty$
- For sex, u, we have just two values: $\tilde{u} \in \{M, F\}$

$$f_S(\tilde{t}_j, \tilde{u}) = K(\tilde{t}_j, \tilde{u})^{-1} \sum_{\mathcal{X}(\tilde{t}_j, \tilde{u})} f(\tilde{t}_j, \tilde{u}, v_i, w_i)$$

where $\mathcal{X}(\tilde{t}_j, \tilde{u}) = \{i : \tilde{t}_{j-1} < t_i < \tilde{t}_{j+1},\ u_i = \tilde{u}\}$
and $K(\tilde{t}_j, \tilde{u}) = |\mathcal{X}(\tilde{t}_j, \tilde{u})|$

MLR marginal effects and computational efficiency

- How can marginal effects be calculated efficiently with BART?
- And beyond BART, many of the ideas that we will explore here can be readily adapted to other MLR methods
- Nearest Neighbor Marginals are generally efficient, but may not be applicable to every problem
- For large training sets, FPD can be computationally demanding whether assuming independence or with Direct Imputation
- In these cases, we are seeking a faster marginal method than FPD
- We can speed up FPD by *random sampling*
  Lundberg and Lee 2017 *NIPS*
  Janzing, Minorics and Blobaum 2020 *PMLR*

FPD vs. FPD by random sampling

FPD

$$f_{S_{F_m}}(x_S) \equiv N^{-1} \sum_i f_m(x_S, x_{iC})$$

where x_{iC} is a training value

$$\hat{f}_{S_F}(x_S) \equiv M^{-1} \sum_m f_{S_{F_m}}(x_S)$$

FPD by random sampling

$$f_{S^K_{F_m}}(x_S) \equiv K^{-1} \sum_k f_m(x_S, x_{k_m C})$$

$x_{k_m C}$ is a draw from the training

$$\hat{f}_{S^K_F}(x_S) \equiv M^{-1} \sum_m f_{S^K_{F_m}}(x_S)$$

FPD by random sampling and the empirical variance

- It is clear that $\mathbf{E}\left[\hat{f}_{S_F}(x_S)\right] \approx \mathbf{E}\left[\hat{f}_{S_F^K}(x_S)\right]$
- However, it is also clear that the variances are not equal

$$\begin{aligned}
\mathbf{V}\left[\hat{f}_{S_F^K}(x_S)|y\right] =&\mathbf{V}\left[\mathbf{E}\left[\hat{f}_{S_F^K}(x_S)|\hat{f}_{S_F}(x_S),y\right]|y\right] \\
&+\mathbf{E}\left[\mathbf{V}\left[\hat{f}_{S_F^K}(x_S)|\hat{f}_{S_F}(x_S),y\right]|y\right] \\
=&\mathbf{V}\left[\hat{f}_{S_F}(x_S)|y\right] \\
&+\mathbf{E}\left[K^{-1}\mathbf{V}\left[f(x_S,x_{kC})|\hat{f}_{S_F}(x_S),y\right]|y\right] \\
\approx&\mathbf{V}\left[\hat{f}_{S_F}(x_S)|y\right]+K^{-1}\mathbf{E}\left[s^2_{S_F^K(x_S)}|y\right]
\end{aligned}$$

$$\text{where } s^2_{S_F^K(x_S)} = K^{-1}\sum_k \left(f(x_S,x_{kC}) - \hat{f}_{S_F^K}(x_S)\right)^2$$

FPD by random sampling and the empirical variance

$$\mathbf{V}\left[\hat{f}_{S_F^K}(x_S)|y\right] \approx \mathbf{V}\left[\hat{f}_{S_F}(x_S)|y\right] + K^{-1}\mathbf{E}\left[s^2_{S_F^K(x_S)}|y\right]$$

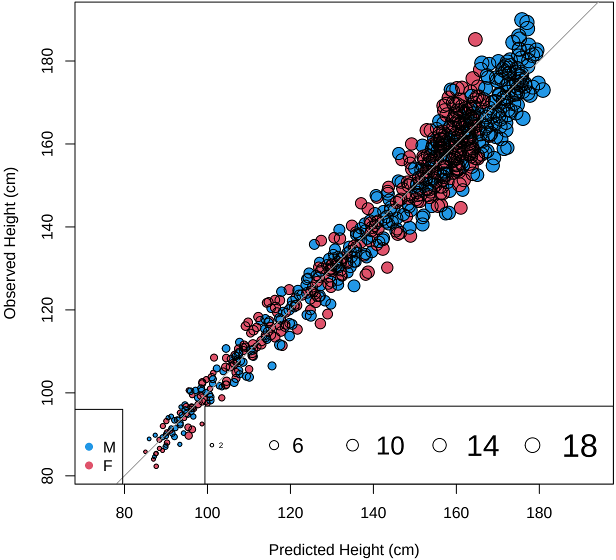
- The first term $\mathbf{V}\left[\hat{f}_{S_F}(x_S)|y\right]$ is the target variance of the calculation we want to avoid
- And the second term can be estimated from the posterior as $\widehat{s^2}_{S_F^K(x_S)} = M^{-1}\sum_m s^2_{S_{F_m}^K(x_S)}$
- Therefore, we can empirically estimate the variance like so $\mathbf{V}\left[\hat{f}_{S_F}(x_S)|y\right] \approx \mathbf{V}\left[\hat{f}_{S_F^K}(x_S)|y\right] - K^{-1}\widehat{s^2}_{S_F^K(x_S)}$
- So, we generate the posterior for the random sampling estimator $f_{S_{F_m}}(x_S) \approx \hat{f}_{S_F^K}(x_S) + \left[f_{S_{F_m}^K}(x_S) - \hat{f}_{S_F^K}(x_S)\right]\sqrt{\frac{\mathbf{V}\left[\hat{f}_{S_F}(x_S)|y\right]}{\mathbf{V}\left[\hat{f}_{S_F^K}(x_S)|y\right]}}$

Returning to the real data example

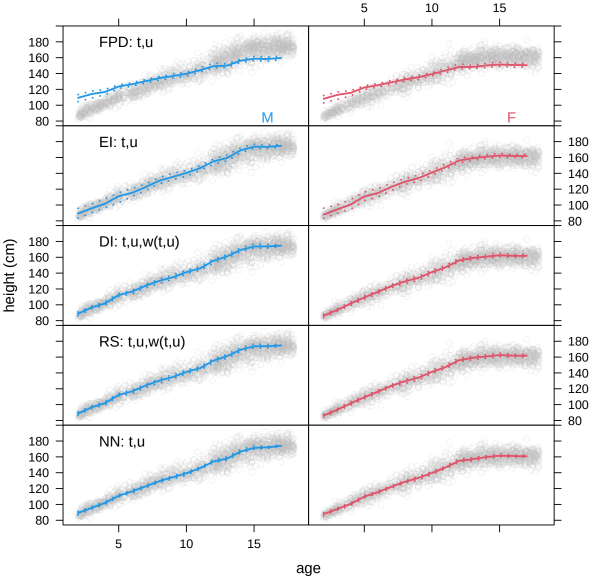
- CDC's data is the US National Health and Nutrition Examination Survey (NHANES) waves I-III
  circa 1972 (I), 1978 (II), 1991 (III): n=12677
- For simplicity, I used NHANES annual/continuous 1999-2000
- The data set is in the BART3 package: `bmx`
  see the `growth*.R` examples in `demo`
- 2-17 years (fractional age for months)
- each child only measured once
- height (cm) and weight (kg) collected
- Check MCMC convergence with **max $\widehat{R}$ < 1.1** for σ:
  Vehtari, Gelman et al. 2021 *Bayesian Analysis*

	n	%
Total	3435	
Males	1768	51.5
Females	1667	48.5
White	800	23.3
Black	1035	30.1
Hispanic	1600	46.6

$R^2 = 96.2\%$ in the testing subset: `growth1.R`



Marginal effect of age comparison



Heteroskedastic BART (HBART)

Pratola, Chipman, George & McCulloch 2020 *JCGS*

$$y_i = \mu_0 + f(x_i) + s(x_i)\epsilon_i \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}\left(0,\ w_i^2\sigma^2\right)$$

$$f \overset{\text{prior}}{\sim} \text{BART}\,(a, b, H, \kappa, \mu_0, \tau)$$

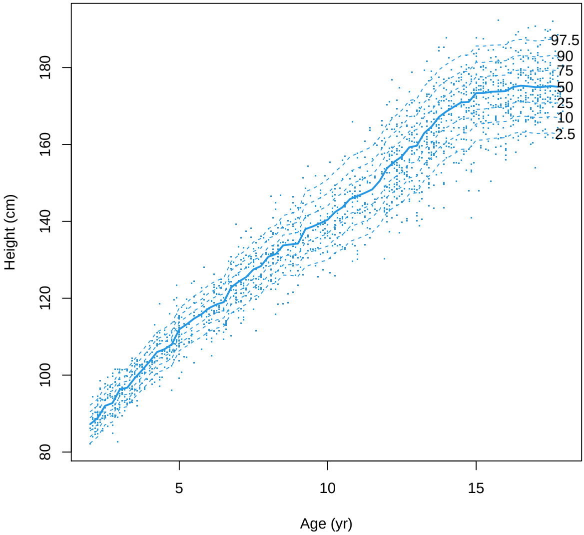
$$s^2 \overset{\text{prior}}{\sim} \text{HBART}\,(\tilde{a}, \tilde{b}, \widetilde{H}, \tilde{\lambda}, \tilde{\nu})$$

$$s^2(x_i) \equiv \prod_{h=1}^{\widetilde{H}} g(x_i; \widetilde{\mathcal{T}}_h, \widetilde{\mathcal{M}}_h) \qquad \widetilde{H} \approx H/5$$

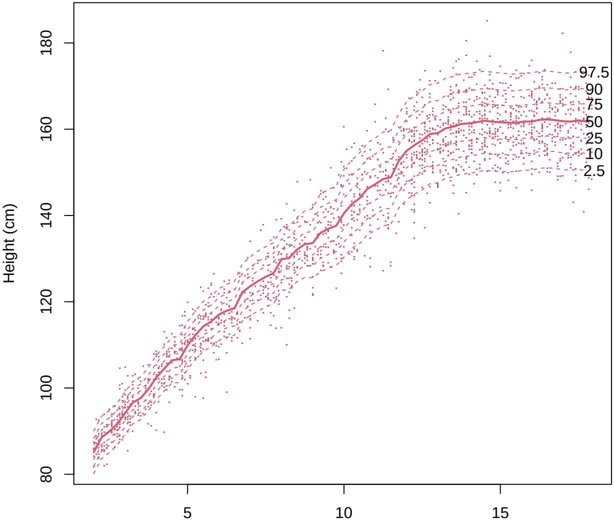
$$\sigma_{hl}^2 | \widetilde{\mathcal{T}}_h \overset{\text{prior}}{\sim} \lambda\nu\chi^{-2}(\nu) \text{ leaves of } \widetilde{\mathcal{T}}_h \qquad \lambda = \tilde{\lambda}^{1/\widetilde{H}}$$

$$\in \widetilde{\mathcal{M}}_h \qquad \nu = 2\left[1 - \left(1 - \frac{2}{\tilde{\nu}}\right)^{1/\widetilde{H}}\right]^{-1}$$

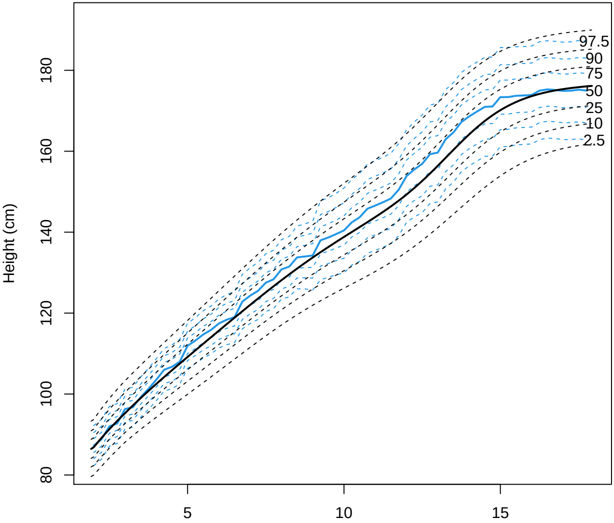
Marginal effect of age: HBART predictions for M
Direct Imputation Marginal: **hbart** `demo/height`



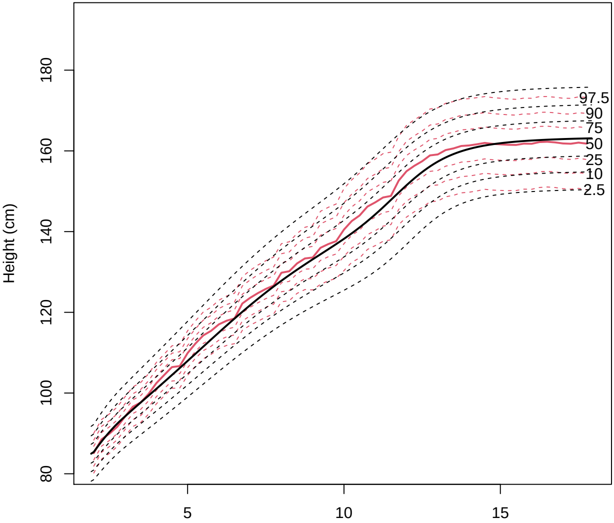
Marginal effect of age: HBART predictions for F
Direct Imputation Marginal: **hbart** `demo/height`



Marginal effect of age: HBART vs. CDC for M
Direct Imputation Marginal: **hbart** `demo/height`



Marginal effect of age: HBART vs. CDC for F
Direct Imputation Marginal: **hbart** `demo/height`



MLR marginal effects and computational efficiency

- *Shapley values* are a popular choice for explainability that are based on marginal effects
- However, Shapley values are very computationally intensive (with their typical naive definition): not a reasonable alternative unless the number of covariates is small
- Shapley values approximate $f(x)$ by additive effects (typically one variable at a time), e.g., $f(x) \approx \sum_j f_j(x_j)$
- $f(x)$ is additive in terms of single covariate functions, $f_j(x_j)$, i.e., effectively, we are assuming independence
- But there is a common extension for two-way interactions Lundberg and Erion et al. 2020 *Nature Machine Intelligence*

Shapley value marginal effects of Independent Covariates

- **Two equivalent definitions: original ordered vs. more computationally friendly unordered**
- $\mathcal{P}_j$ is the set of all *ordered* permutations of $C_{-j} \cup \{x_j\}$
  $f_j(x_j) \equiv (P!)^{-1} \sum_{S_* \in \mathcal{P}_j} [f_j^*(x_{S_*}) - f_{-j}^*(x_{S_*})]$
  where $f_j^*(x_{S_*})$ only evaluates arguments up to/including x_j
  and $f_{-j}^*(x_{S_*})$ only evaluates arguments before/excluding x_j
- C_j^* is the set of all *unordered* combinations $C_* \subset C_{-j}$
  $f_j(x_j) \equiv \sum_{C_* \in C_j^*} \frac{|C_*|!(P-|C_*|-1)!}{P!} [f_{S^j}(x_j, x_{C_*}) - f_{S^{-j}}(x_{C_*})]$
- If each $f_S(.)$ are fit from the training
  the number of fits needed grows rapidly with P

P	2	3	4	5	10	20	30	P
Fits	3	7	15	31	1,023	1,048,575	1,073,741,823	$2^P - 1$

Fast Shapley value approximations from a single fit

- Rather than fitting so many models, Shapley values can be created from a single fit's marginal effects
- For example, suppose $f_S(x_S) = \mathrm{E}_{x_{C_*}}\left[f(x_S, x_{C_*})|x_S\right]$
- This would certainly help but the computations are still daunting unless the number of covariates is small
- There is a simple algorithm, `EXPVALUE`, for these marginals that is basically equivalent to FPD
  And there are more efficient, so-called Tree SHAP, algorithms but these are far more complex
  Lundberg and Erion et al. 2020 *Nature Machine Intelligence*
- And advanced random sampling schemes have been proposed but they are challenging to implement as well
  Yang, Zhou et al. 2023 *JASA*

Shapley value marginal effects of Dependent Covariates
Marginal effect of age

- Shapley values come from game theory where
  each player takes their turn and the order of play is important
- The *players* here are the covariates
- And as can be shown, the order of covariates doesn't really matter
  i.e., the order of covariates is arbitrary (Lundberg and Lee 2017)
- Nevertheless, all possible orderings of $\boldsymbol{t, u, v, w}$: $\boldsymbol{P! = 24}$

age first	age second	age third	age last
$\boldsymbol{t, u, v, w}$	$\boldsymbol{u, t, v, w}$	$\boldsymbol{u, v, t, w}$	$\boldsymbol{u, v, w, t}$
$\boldsymbol{t, u, w, v}$	$\boldsymbol{u, t, w, v}$	$\boldsymbol{u, w, t, v}$	$\boldsymbol{u, w, v, t}$
$\boldsymbol{t, v, u, w}$	$\boldsymbol{v, t, u, w}$	$\boldsymbol{v, u, t, w}$	$\boldsymbol{v, u, w, t}$
$\boldsymbol{t, v, w, u}$	$\boldsymbol{v, t, w, u}$	$\boldsymbol{v, w, t, u}$	$\boldsymbol{v, w, u, t}$
$\boldsymbol{t, w, u, v}$	$\boldsymbol{w, t, u, v}$	$\boldsymbol{w, u, t, v}$	$\boldsymbol{w, u, v, t}$
$\boldsymbol{t, w, v, u}$	$\boldsymbol{w, t, v, u}$	$\boldsymbol{w, v, t, u}$	$\boldsymbol{w, v, u, t}$

Shapley value marginal effects of Dependent Covariates
Marginal effect of age

Differentials for $\boldsymbol{t}$ corresponding to each ordering

$f(t)-0$	$f(u,t)-f(u)$	$f(u,v,t)-f(u,v)$	$f(u,v,w,t)-f(u,v,w)$
$f(t)-0$	$f(u,t)-f(u)$	$f(u,w,t)-f(u,w)$	$f(u,w,v,t)-f(u,w,v)$
$f(t)-0$	$f(v,t)-f(v)$	$f(v,u,t)-f(v,u)$	$f(v,u,w,t)-f(v,u,w)$
$f(t)-0$	$f(v,t)-f(v)$	$f(v,w,t)-f(v,w)$	$f(v,w,u,t)-f(v,w,u)$
$f(t)-0$	$f(w,t)-f(w)$	$f(w,u,t)-f(w,u)$	$f(w,u,v,t)-f(w,u,v)$
$f(t)-0$	$f(w,t)-f(w)$	$f(w,v,t)-f(w,v)$	$f(w,v,u,t)-f(w,v,u)$

Weighted differentials for $\boldsymbol{t}$ corresponding to each ordering

$6f(t)$	$2[f(t,u)-f(u)]$	$2[f(t,u,v)-f(u,v)]$	$6[f(t,u,v,w)-f(u,v,w)]$
	$2[f(t,v)-f(v)]$	$2[f(t,u,w)-f(u,w)]$	
	$2[f(t,w)-f(w)]$	$2[f(t,v,w)-f(v,w)]$	
0	1	2	3
3!	**2!**	**2!**	**3!**

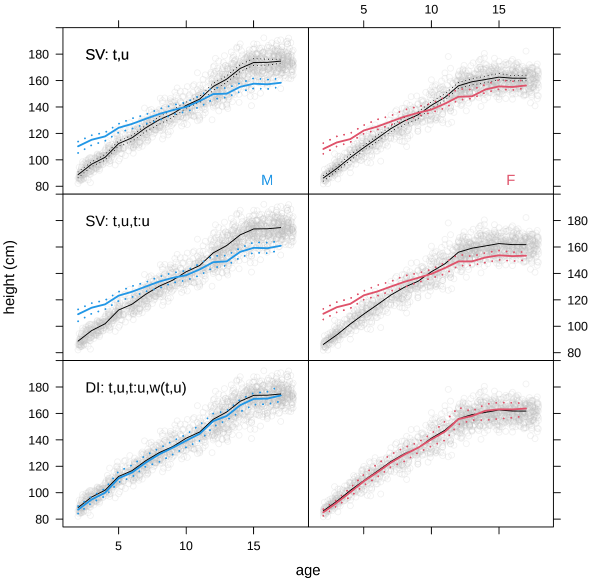
Last row are the weights for the differentials: $|C_*|!(P-|S|-|C_*|)!$

(Lundberg and Lee 2017)

Shapley values and
Marginal Effects for Dependent Covariates
Extending Direct Imputation Marginal to SHAP

- As before, rely on the strong relationships of age, sex and weight
  $\mathbf{E}[w|t,u] = \tilde{w} = \tilde{\mu} + \tilde{f}(t,u)$
  where $w_i = \tilde{\mu} + \tilde{f}(t_i,u_i) + \tilde{\epsilon}_i$ **where** $\tilde{f} \overset{\text{prior}}{\sim}$ **BART**
- For a marginal effect more applicable to dependent variables
  Females: $f_t(t) + f_u(F) + 2f_{t:u}(t,F) + f_w(\tilde{\mu} + \tilde{f}(t,F))$

Marginal effects with Shapley values



Marginal effect of age: computational efficiency measured by `system.time()` in seconds

	Computational Timings			
	`user`		`elapsed`	
Method	s	%	s	%
FPD: Direct Imputation Marginal	340	100	64	100
FPD: Nearest Neighbor Marginal	32	9	20	31
FPD: Random Sampling $\boldsymbol{K = 30}$	130	38	17	27
FPD: Random Sampling $\boldsymbol{K = 5}$	22	6	3	5
SHAP: $\boldsymbol{t}$, age-only	1610		1610	
SHAP: $\boldsymbol{u}$, sex-only	249		249	
SHAP: $\boldsymbol{w}$, weight-only	2007		2011	
SHAP: Direct Imputation Marginal	3866	1137	3870	6047

Marginal effects for dependent covariates and computational efficiency

- At first, it is quite surprising that FPD assumes independence since it has the term *dependence* in its name
- Our new proposed marginals Nearest Neighbor and Random Sampling with Direct Imputation are computationally efficient
- But the Shapley value marginals are very computationally demanding and impractical
- It is possible to exploit the structure of binary trees to compute Shapley values by the so-called Tree SHAP algorithms Lundberg and Erion et al. 2020 *Nature Machine Intelligence* for example, see the **treeshap** R package for Random Forests but whether that makes them feasible is not yet known
- My **BART3** package on `github` has S3 methods for `FPD/SHAP` and their countparts with random sampling

Conclusion

- This was an overview of BART and its place in machine learning
- Our focus was on the BART prior for continuous outcomes
- In particular, estimating marginal effects with BART whether assuming independence or dependence
- We contrasted Friedman's partial dependence function with Shapley values
- And we have described facilitating these calculations with opportunities for bettering performance statistically and computationally
- We provide a reference implementation in the **BART3** R package with *new and improved* marginal effects S3 functions