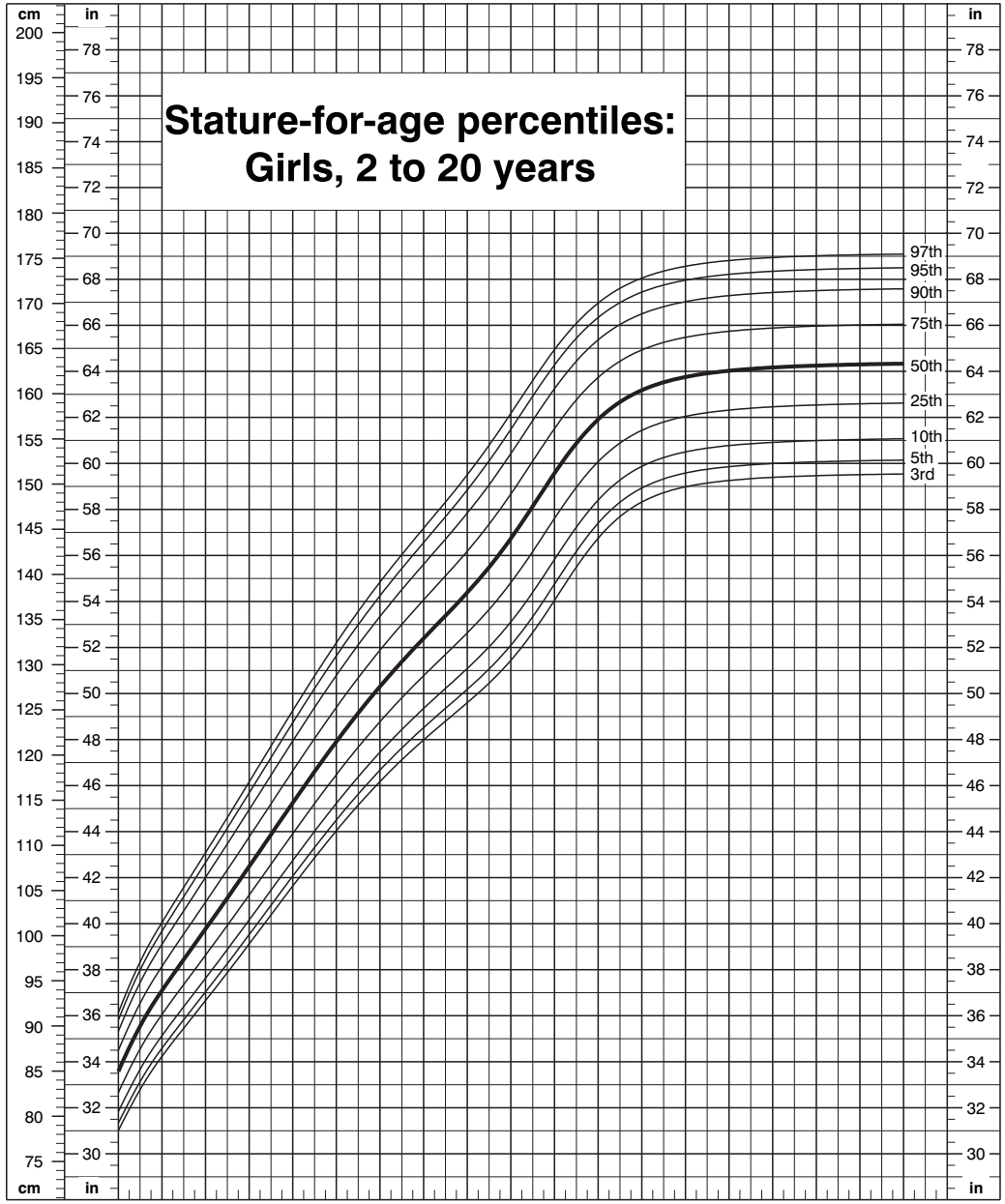
Machine learning regression and marginal effects inference

Rodney Sparapani
Associate Professor of Biostatistics
Medical College of Wisconsin

September 15, 2025

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

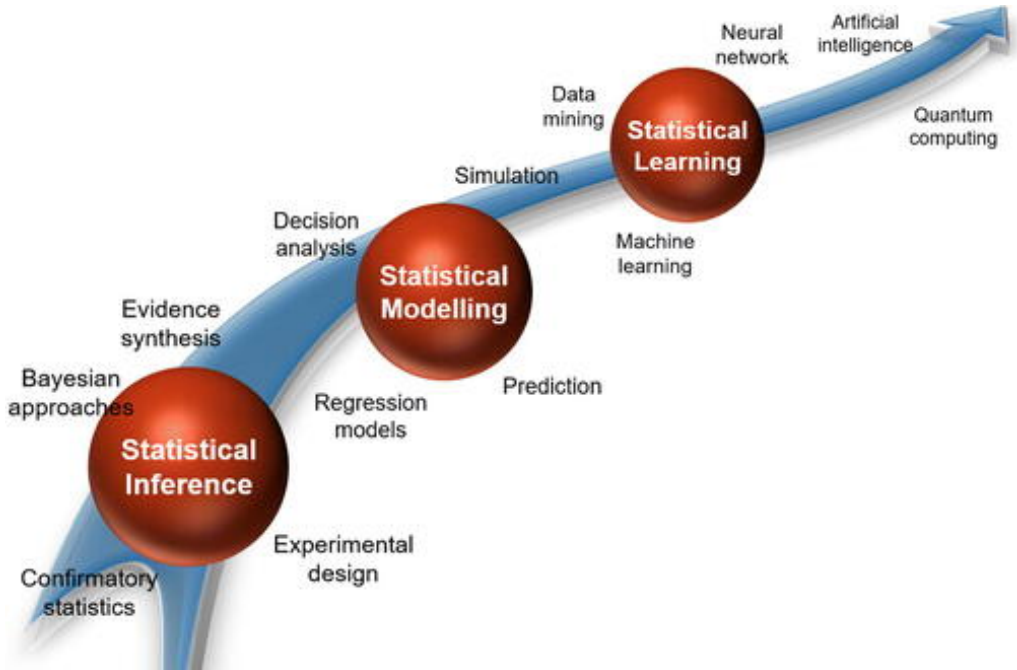
$$z_t = \left\{ \begin{array}{ll} \frac{-1+(y_t/M_t)^{L_t}}{L_t S_t} & L_t \neq 0 \\ \frac{\log(y_t/M_t)}{S_t} & L_t = 0 \end{array} \right\} \sim \mathrm{N}(0,\ 1)$$

- 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 let's not bother even 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 we focus 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

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

- f is an unspecified function whose form is to be *learned* from the training data and $\boldsymbol{x_i}$ is a vector of covariates for $\boldsymbol{i = 1, \ldots, N}$ following Samuel's principle via Bayesian nonparametric models without resorting to precarious restrictive assumptions, i.e., we don't want to assume linearity nor pre-specify interactions, etc.

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"

Single-tree regression model

Chipman, George & McCulloch 1998 *JASA*

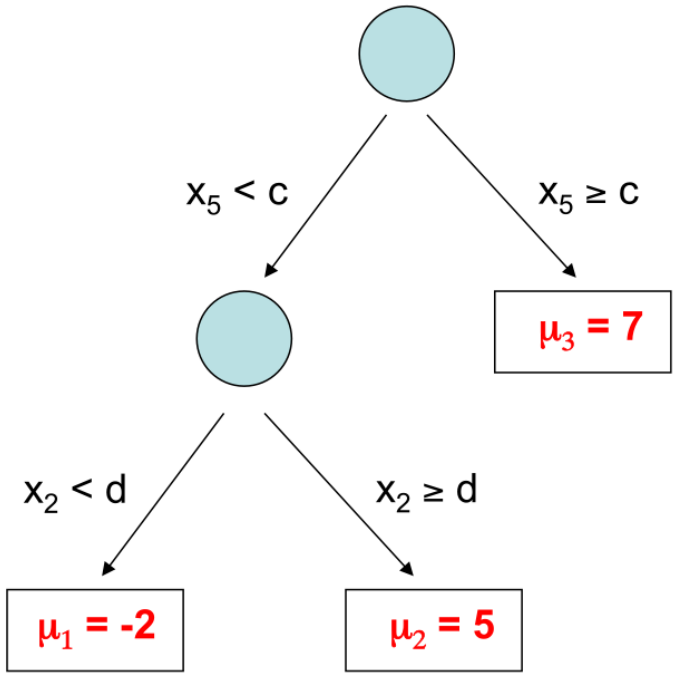
y_i is a continuous outcome where i indexes subjects $i = 1, \ldots, N$

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*

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

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

$$f(x_i) \equiv \mu_0 + \sum_{h=1}^{H} g(x_i; \mathcal{T}_h, \mathcal{M}_h) \qquad H \in \{50, 200, 500\}$$

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

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

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

The BART 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	$\mathbf{P[branch}|\text{depth}] = \alpha(1+\text{depth})^{-\beta}$
Default prior settings $\alpha = 0.95, \beta = 2$	

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

Marginal Effects and Machine Learning Regression (MLR)

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

$$\mathbf{E}[y|\boldsymbol{x}_S] \equiv f_S(\boldsymbol{x}_S)$$

$$f_S(\boldsymbol{x}_S) \equiv \mathbf{E}_{\boldsymbol{x}_C}[f(\boldsymbol{x}_S, \boldsymbol{x}_C)|\boldsymbol{x}_S]$$

$$= \int \cdots \int f(\boldsymbol{x}_S, \boldsymbol{x}_C)\,[\boldsymbol{x}_C|\boldsymbol{x}_S]\,\mathrm{d}\boldsymbol{x}_C$$

where $[\boldsymbol{x}_C|\boldsymbol{x}_S]$ is the distribution of $\boldsymbol{x}_C|\boldsymbol{x}_S$

$$\approx \int \cdots \int f(\boldsymbol{x}_S, \boldsymbol{x}_C)\,[\boldsymbol{x}_C]\,\mathrm{d}\boldsymbol{x}_C \qquad \text{assuming } \boldsymbol{x}_S \perp \boldsymbol{x}_C$$

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

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

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

$$\approx N^{-1} \sum_i f(x_S, x_{iC})$$

the partial dependence function

where x_{iC} are the training values

$$f_{Sm}(x_S) \approx N^{-1} \sum_i f_m(x_S, x_{iC})$$

for each MCMC sample m

$$\hat{f}_S(x_S) \approx M^{-1} \sum_m f_{S_m}(x_S)$$

estimate by averaging over the posterior

Friedman 2001 *Annals of Statistics*

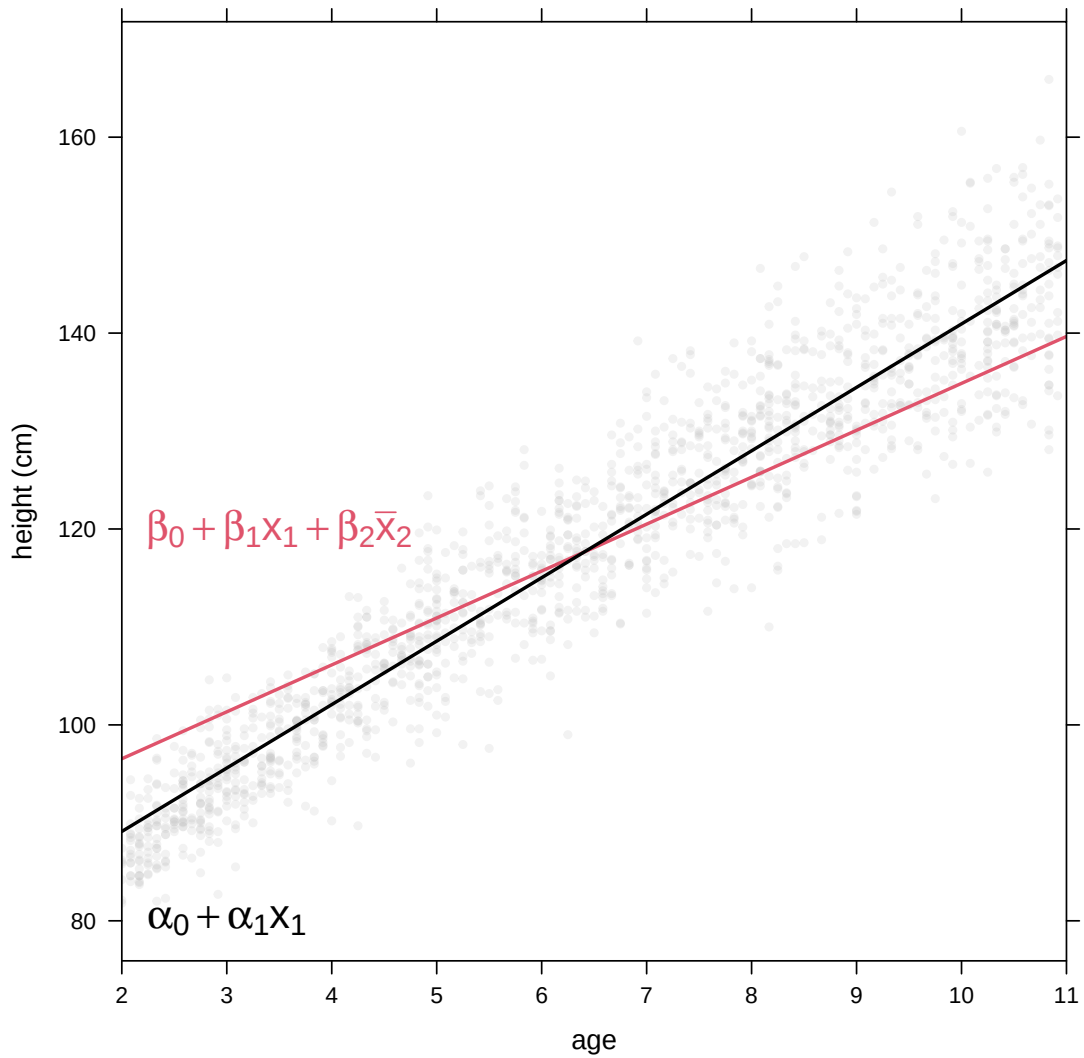
Friedman's partial dependence function (FPD) and Marginal Effects Assuming Independent Covariates

Linear regression example: age $= x_S = x_1$, weight $= x_C = x_2$

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

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

Linear regression example: age $= x_S = x_1$, weight $= x_C = x_2$



Independent Covariates: Simple Random Sampling (SRS) marginal

Lundberg & Lee 2017 *NIPS*

Janzing et al. 2020 *PMLR*

- To speed up FPD when N is large
- Consider our growth chart for height example
  age $= t$, sex $= u$, race $= v$ and weight $= w$
  $y_i = f(t_i, u_i, v_i, w_i) + \epsilon_i$ **where** $f \overset{\text{prior}}{\sim}$ **BART**
- $S = (t, u)$ and $C = (v, w)$
- Like FPD, but with K random draws of (v_i, w_i) from training
- K does not have to be large, e.g., $K = 30$ is often sufficient

$$f_{t,u}(t,u) = \mathrm{E}_{v,w}\left[f(t,u,v,w)|t,u\right]$$
$$\approx K^{-1}\sum_k f(t,u,v_k^*,w_k^*)$$

where $(v_k^*, w_k^*) \sim \{(v_1, w_1), \ldots, (v_N, w_N)\}$ simple random sample

Dependent Covariates: the Monte Carlo (MC) marginal

- Consider our growth chart for height example
  age $= t$, sex $= u$, race $= v$ and weight $= w$
  $y_i = f(t_i, u_i, v_i, w_i) + \epsilon_i$ **where** $f \overset{\text{prior}}{\sim}$ **BART**
- Next we model the strong relationship between covariates
  $\mathbf{E}[w|t, u, v] = \tilde{w} = \tilde{f}(t, u, v)$
- Here we summarize this relationship with a BART model
  $w_i = \tilde{f}(t_i, u_i, v_i) + \tilde{\epsilon}_i$ **where** $\tilde{f} \overset{\text{prior}}{\sim}$ **BART**
- For marginal effects applicable to dependent variables
- $S = (t, u)$ and $C = (v, w)$

$$f_{t,u}(t, u) = \mathbf{E}_{v,w}[f(t, u, v, w)|t, u]$$
$$\approx K^{-1} \sum_k f(t, u, v_k^*, w_k^*)$$

where $v_k^* \sim \{v_1, \ldots, v_N\}$ simple random sample

and $w_k^* \sim \mathrm{N}\left(\tilde{f}(t, u, v_k^*),\ \sigma_{\tilde{\epsilon}}^2\right)$

Dependent Covariates: a hybrid marginal

- Consider our growth chart for height example
  age $= t$, sex $= u$, race $= v$ and weight $= w$
  $y_i = f(t_i, u_i, v_i, w_i) + \epsilon_i$ **where** $f \overset{\text{prior}}{\sim}$ **BART**
- Next we model the strong relationship between covariates
  $\mathbf{E}[w|t,u] = \tilde{w} = \tilde{f}(t,u)$
- Here we summarize this relationship with a BART model
  $w_i = \tilde{f}(t_i, u_i) + \tilde{\epsilon}_i$ **where** $\tilde{f} \overset{\text{prior}}{\sim}$ **BART**
- For marginal effects applicable to dependent variables
- $S = (t,u)$ and $C = (v,w)$

$$f_{t,u}(t,u) = \mathbf{E}_{v,w}[f(t,u,v,w)|t,u]$$
$$\approx K^{-1} \sum_k f(t,u,v_k^*, w_k^*)$$

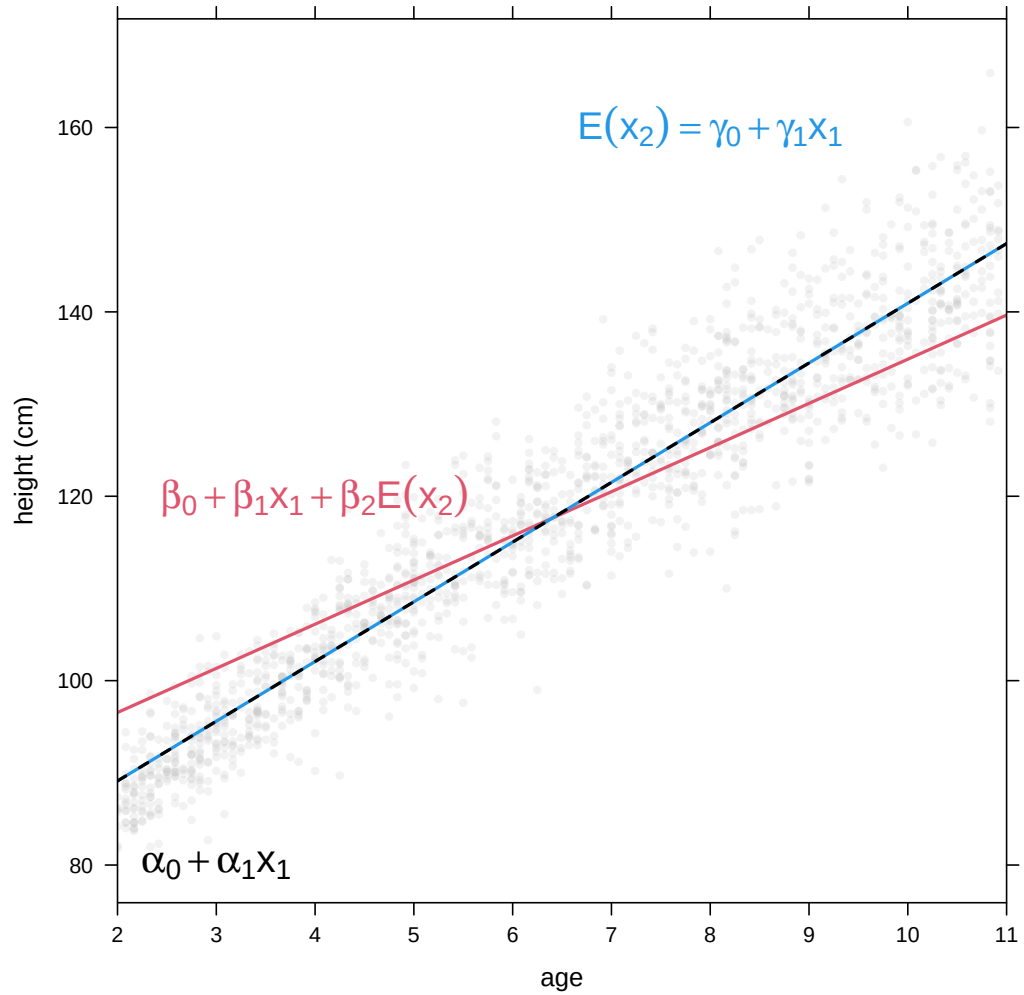
where $v_k^* \sim \{v_1, \ldots, v_N\}$ simple random sample

and $w_k^* \sim \mathrm{N}\left(\tilde{f}(t,u),\ \sigma_{\tilde{\epsilon}}^2\right)$

Dependent Covariates: the Synthetic Approximation (SA) marginal

Linear regression example: age $= x_S = x_1$, weight $= x_C = x_2$

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



Dependent Covariates: the Synthetic Approximation (SA) marginal

- We proceed as before by modeling the strong relationship between age, sex and weight first
- For marginal effects applicable to dependent variables
- $S = (t, u)$ and $C = (v, w)$

$$
\begin{aligned}
\mathbf{E}[w|t,u] &= \tilde{f}(t,u) \\
w_i &= \tilde{f}(t_i,u_i) + \tilde{\epsilon}_i \text{ where } \tilde{f} \overset{\text{prior}}{\sim} \text{BART} \\
\widehat{w}(t,u) &= M^{-1} \sum_m \tilde{f}_m(t,u) \\
f_{t,u}(t,u) &= \mathbf{E}_v[f(t,u,v,w)|t,u,w = \mathbf{E}[w|t,u]] \\
&= \mathbf{E}_v[f(t,u,v,\widehat{w}(t,u))|t,u] \\
&\approx N^{-1} \sum_i f(t,u,v_i,\widehat{w}(t,u))
\end{aligned}
$$

Dependent Covariates:
the Nearest Neighbor (NN) marginal

- t for age, u for sex, v for race/ethnicity and w for weight
- For age, t, we have chosen this grid of values
  $-\infty = t_0^* < t_1^* < t_2^* < \cdots < t_J^* < t_{J+1}^* = \infty$
  where $t_1^* = 2, \ldots, t_{17}^* = 18$
- For sex, u, we have just two values: $u^* \in \{M, F\}$

$$f_S(t_j^*, u^*) \approx K(t_j^*, u^*)^{-1} \sum_{\mathcal{X}(t_j^*, u^*)} f(t_j^*, u^*, v_i, w_i)$$

where $\mathcal{X}(t_j^*, u^*) = \{i : t_{j-1}^* < t_i < t_{j+1}^*,\ u_i = u^*\}$

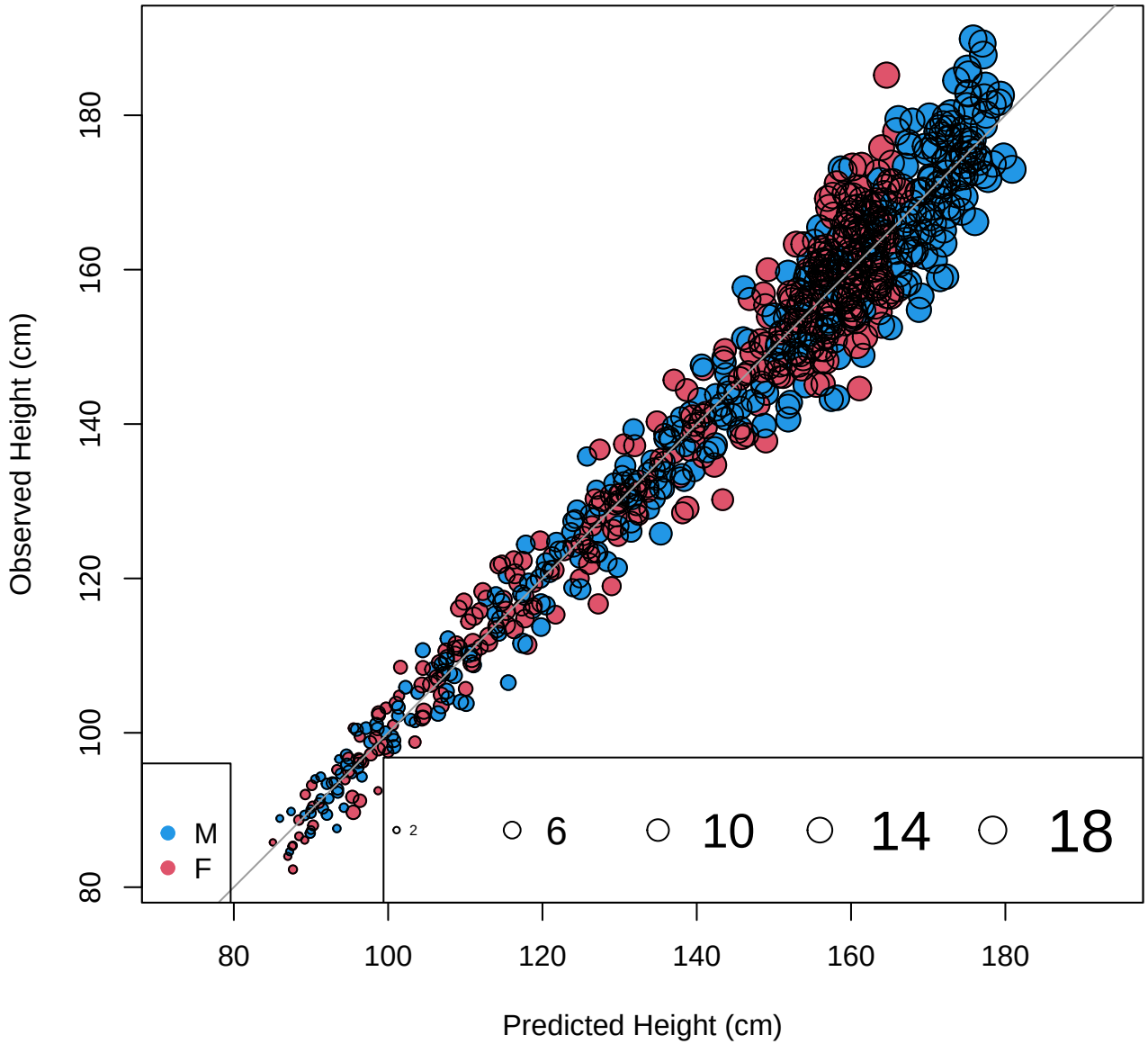
and $K(t_j^*, u^*) = |\mathcal{X}(t_j^*, u^*)|$

Returning to the real data example

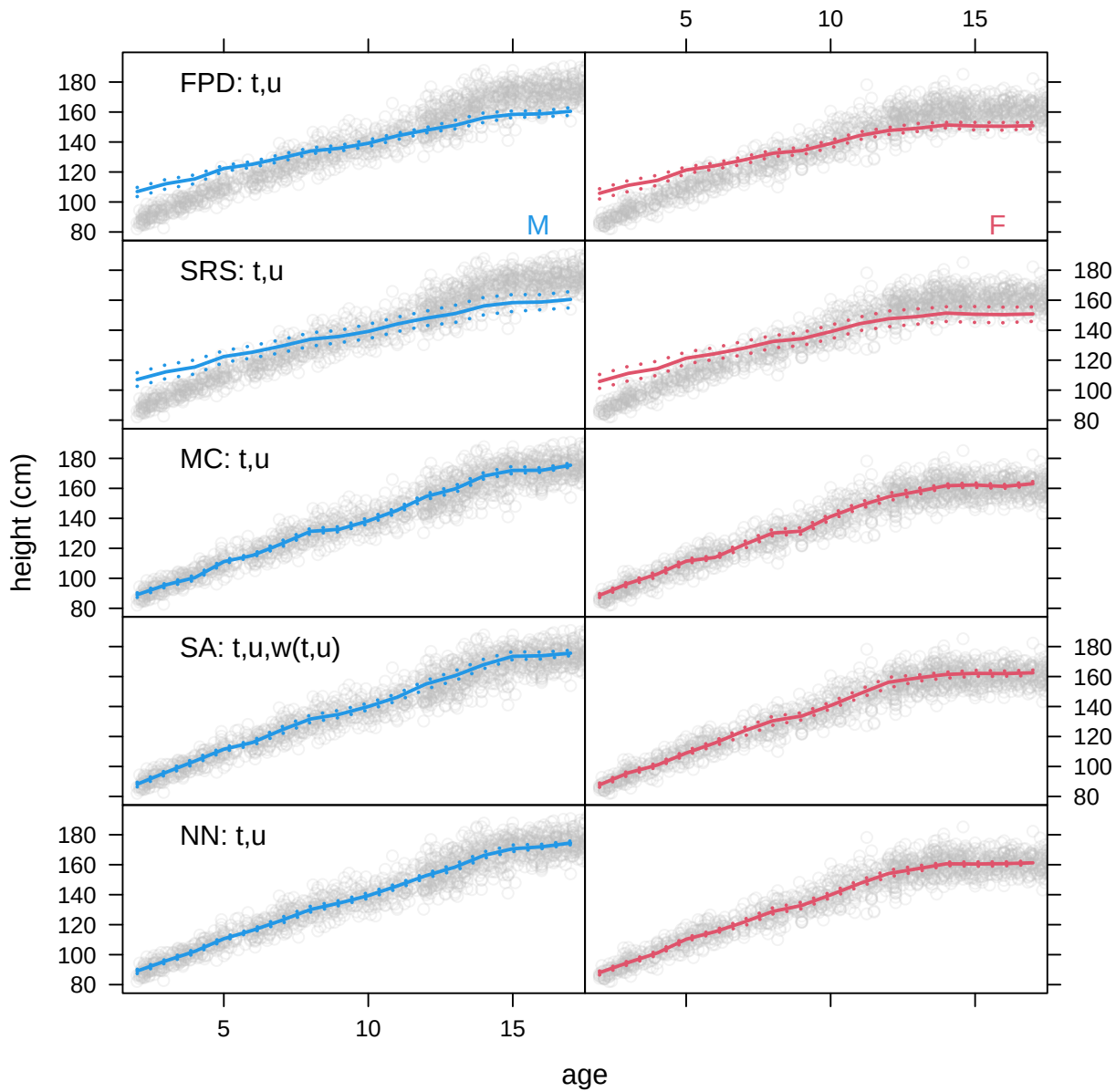
- CDC's data is the US National Health and Nutrition Examination Survey (NHANES) waves I-III
  circa 1972 (I), 1978 (II), 1991 (III): ***n***=12677
- For simplicity, I used NHANES annual/continuous 1999-2000
- The data set is in the BART3 package: `bmx`
  see the `growth*.R` examples in `demo`
- 2-17 years (fractional age for months)
- each child only measured once
- height (cm) and weight (kg) collected

	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



MLR marginal effects and computational efficiency

- *Shapley values* are a popular choice for explainability research that are based on marginal effects
- However, Shapley values are very computationally intensive
  In terms of complexity, they are considered to be NP-hard:
  not practical unless the number of covariates is small
- Shapley values approximate $f(x)$ by additive effects (typically of one variable at a time), e.g., $f(x) \approx \mu + \sum_j f_j(x_j)$
- $f(x)$ 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*
- And multi-way interactions have been defined as well
  Grabisch & Roubens 1999; Borgonovo et al. 2024

Shapley value marginal effects of Independent Covariates

- **Two equivalent definitions: original ordered vs. more computationally friendly unordered**
- $S = \{x_j\}$ and C_{-j} contains all $P - 1$ other covariates
- $\mathcal{P}_j$ is the set of all *ordered* permutations of $S \cup C_{-j}$
  $f_j(x_j) \equiv (P!)^{-1} \sum_{S_* \in \mathcal{P}_j} [f^*_{+j}(x_{S_*}) - f^*_{-j}(x_{S_*})]$
  where $f^*_{+j}(x_{S_*})$ evaluates arguments up to/including x_j
  while $f^*_{-j}(x_{S_*})$ evaluates arguments before/excluding x_j
- C_j is the set of all *unordered* combinations $S_* \subset C_{-j}$
  $f_j(x_j) \equiv \sum_{S_* \in C_j} \frac{|S_*|!(P-|S_*|-1)!}{P!} [f_S(x_{S_*}, x_j) - f_S(x_{S_*})]$
- 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, known as `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 to program
  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 marginals for 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 t, u, v, w: $\boldsymbol{P! = 24}$

age first	age second	age third	age last
t, u, v, w	u, t, v, w	u, v, t, w	u, v, w, t
t, u, w, v	u, t, w, v	u, w, t, v	u, w, v, t
t, v, u, w	v, t, u, w	v, u, t, w	v, u, w, t
t, v, w, u	v, t, w, u	v, w, t, u	v, w, u, t
t, w, u, v	w, t, u, v	w, u, t, v	w, u, v, t
t, w, v, u	w, t, v, u	w, v, t, u	w, v, u, t

Shapley value marginal effects of Dependent Covariates
Marginal effect of age

Differentials for 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 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!**

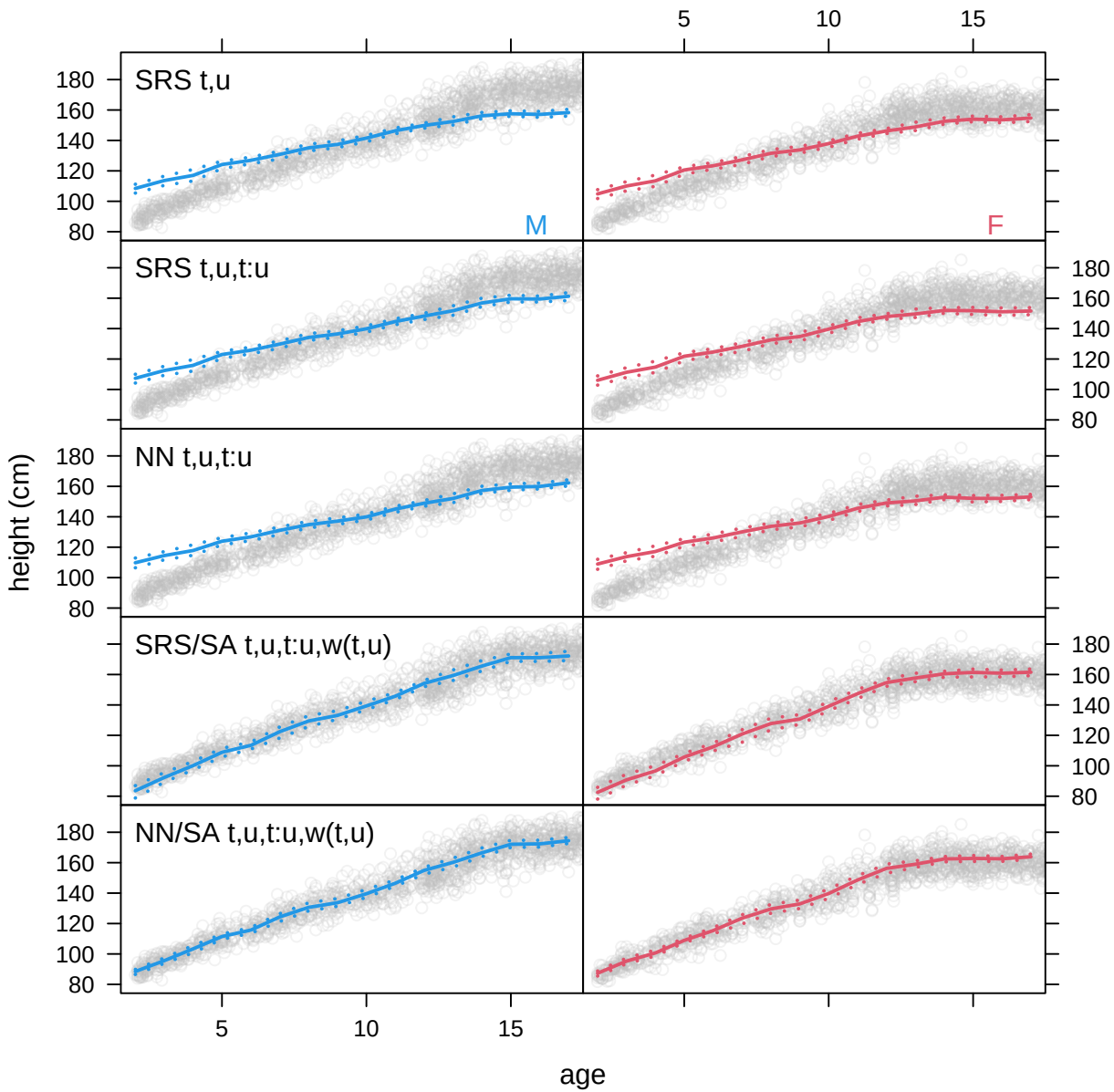
Last row are the weights for the differentials: $\mathbf{|S_*|!(P-|S_*|-1)!}$

(Lundberg and Lee 2017)

Shapley values and
Marginal Effects for Dependent Covariates:
Synthetic Approximation marginal

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

Marginal effects with Shapley values



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

	Computational Timings			
	user		elapsed	
Method	s	%	s	%
Synthetic approximation	340	100	64	100
Nearest neighbor	32	9	20	31
SRS: $\boldsymbol{K = 30}$	130	38	17	27
SRS: $\boldsymbol{K = 5}$	22	6	3	5
SV: $\boldsymbol{t}$, age-only	1610		1610	
SV: $\boldsymbol{u}$, sex-only	249		249	
SV: $\boldsymbol{w}$, weight-only	2007		2011	
SV: Synthetic approximation	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 novel marginals Nearest Neighbor and Synthetic Approximation are computationally efficient
- Yet the Shapley value marginals are very computationally demanding and often 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 this still might not me fast enough to be feasible
- Furthermore, this has not been adapted to BART FWIW

Marginal effects for dependent covariates and computational efficiency

- My **BART3** package on `github` has S3 methods for marginals

S3 method	Assumes $\perp\!\!\!\perp$	Marginal type
`FPD`	Yes	FPD
`SRS`	Yes	Simple random sampling
`SHAP`	Yes	SV
`SHAP2`	Yes	SV two-way interaction
`NN`	No	Nearest neighbors
`SHNN`	No	SV with nearest neighbors
`SHNN2`	No	SV two-way interaction with nearest neighbors

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