BART Marginal Effect Estimation: Efficiency and Extrapolation

Rodney Sparapani
Associate Professor of Biostatistics
Medical College of Wisconsin

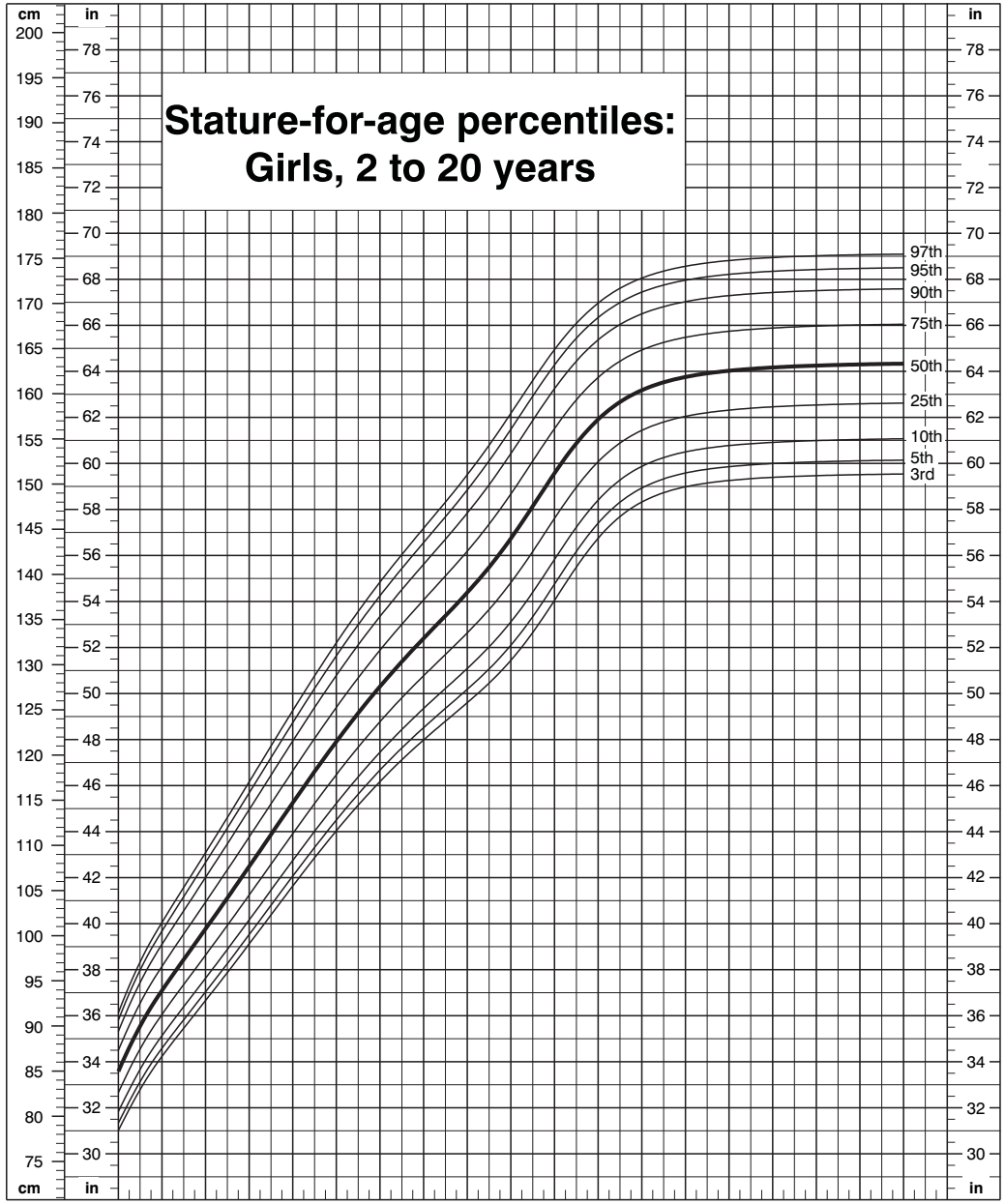
Ensemble Learning with Bayesian Additive Regression Trees
June 7-8, 2023

Funding for this research was provided, in part, by the Advancing Healthier Wisconsin Research & Education Program under award 9520364 and the Children's Wisconsin Foundation through the generous support of the LAMBS fund and Abraham Family Fund

Abstract I: Bayesian Ensemble Learning

Modern computing power has invigorated our ability to learn high-dimensional, complex relationships from data; in particular, two recent breakthroughs: deep learning and ensemble learning. In this workshop, we explore the latter approach via Bayesian ensembles of trees called Bayesian Additive Regression Trees (BART). The Bayesian paradigm naturally provides a Markov chain Monte Carlo stochastic exploration of the model space, uncertainty quantification, and posterior inference. BART is one of the few modern approaches which is able to exploit the powerful Bayesian conceptual toolkit.

CDC Growth Charts: United States



Motivating Example: Growth Charts

- The US Centers for Disease Control and Prevention (CDC) as well as 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)
- The 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/or race/ethnicity?

What is Machine Learning?

- *Artificial intelligence* (AI) is a computer's ability to perform tasks that normally require human intelligence like driving a car
- *Machine learning*, or statistical learning, is a field within AI to develop methods that *learn* predictions from training data *without being explicitly programmed to do so* (paraphrasing Arthur Samuel 1959)
- For example, you could directly model childhood growth chart data based on principles of human auxology or you could indirectly learn the growth curves from training data

What is Machine Learning?

- *Deep learning* is the best currently-known machine learning method of prediction where all of the covariates are of the same type, i.e., they are all pixels or words or audio waves, etc.
- *Ensemble learning* is the best currently-known machine learning method with respect to out-of-sample predictive performance for tabular data where all of the covariates are of different types, i.e., age, sex, height, weight, etc.
  A collection of *machines* (in our case trees) are fit simultaneously that form the basis of an ensemble's aggregate prediction with superior performance to any single machine's fit

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
- Ensembles are in the middle: medium bias/medium variance
- BART is in the class of ensemble models which both theoretically, and in practice, have excellent out-of-sample predictive performance

Krogh & Solich 1997 *Physical Review E*
Baldi & Brunak 2001 "Bioinformatics: machine learning approach"
Kuhn & Johnson 2013 "Applied Predictive Modeling"

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

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

- f is an unspecified function whose form is to be *learned* from the data and x_i is a vector of covariates for $i = 1, \ldots, N$
- An MLR extension we will be discussing today and tomorrow

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

- And f alone (or f and s) will be *learned*, but how?
- *Ideally* in a *nonparametric* manner without resorting to *precarious restrictive assumptions*, i.e., we don't need to assume linearity nor pre-specify interactions

What is Bayesian Additive Regression Trees?

- a supervised MLR with nice properties: automated learning of the functional relationship and interactions without requiring covariate transformations for continuous, binary, categorical and time-to-event outcomes
- tree-based ensemble predictive model
- Bayesian nonparametric method with robust defaults for the prior parameter settings
- computationally efficient posterior inference via MCMC estimates naturally computed from summaries of the posterior along with the quantification of their uncertainty
- seamless extension to variable selection in high dimensions

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*

Selected BART references with URLs: Rob & Rodney

Overview	Chipman, George and McCulloch 2010 *AOAS* Sparapani, Spanbauer and McCulloch 2021 *JSS*
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*
Big Data (Big N)	Pratola, Chipman et al. 2014 *JCGS* Entezari, Craiu et al. 2017 *Canadian J of Stat*
Variable Selection (Big P)	Linero 2018 *JASA* Liu, Rockova 2023 *JASA*
Efficient MCMC	Pratola 2016 *Bayesian Analysis*
Nonparametric Theory	Rockova and Saha 2019 *PMLR* Rockova and van der Pas 2020 *AOS*
Heteroskedastic	Pratola, Chipman et al. 2020 *JCGS*
Propensity Scores	Hahn, Murray et al. 2020 *Bayesian Analysis*
Monotonic	Chipman, George et al. 2021 *Bayesian Analysis*

Bayesian Additive Regression Trees (BART)

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

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

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

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

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

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

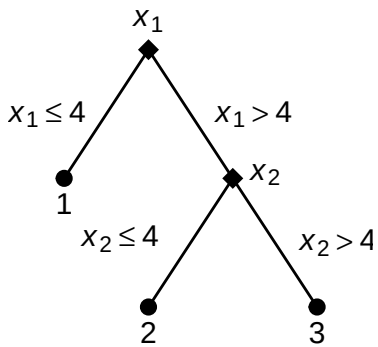
$$\sigma^2 \overset{\text{prior}}{\sim} \lambda\nu\chi^{-2}(\nu)$$

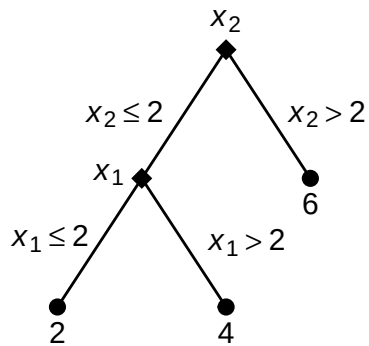
An aside: MLR, BART and careless notation

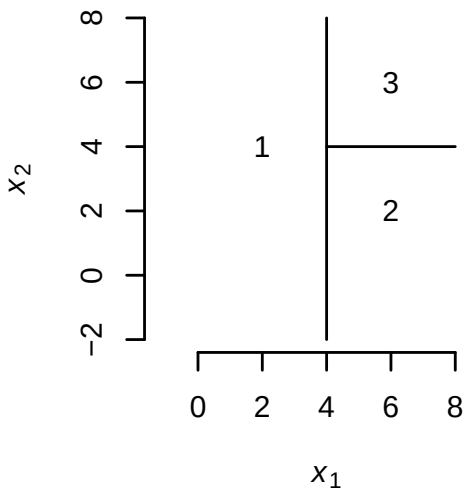
- An important subtlety of MLR/BART notation that is the most common pitfall of the literature/software
- Often authors make the mistake of denoting $f(x)$ when they really mean $\mu + f(x)$
- Rob and I try to avoid this but it is a very easy mistake to make
- Similarly, virtually all MLR/BART software returns $\mu + f(x)$ while not properly documenting it (we are guilty of this as well)
- This is already bad: yet even worse for marginal effects
- Perhaps, we should adopt a new notation like $\mu(x) = \mu + f(x)$ to make the proper distinction between $f(x)$ and $\mu(x)$
- But, that doesn't help with what has already been published
- So, for the most part, we stick to $f(x)$ and $\mu + f(x)$ accordingly

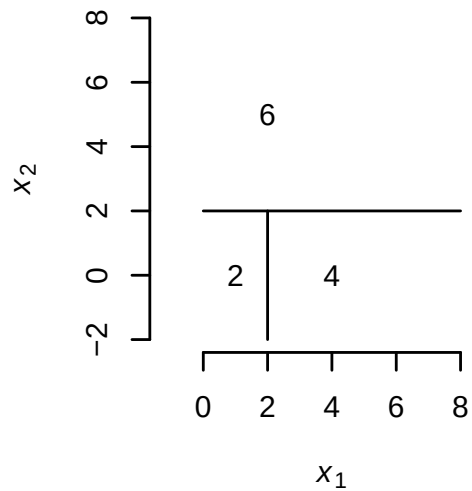
Bayesian Additive Regression Trees (BART)

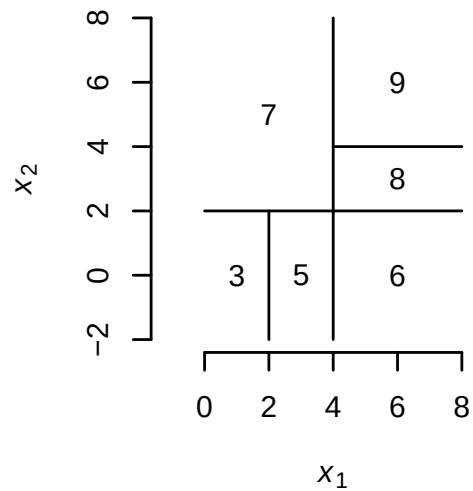
Logan, Sparapani, McCulloch & Laud 2020 *SMMR*









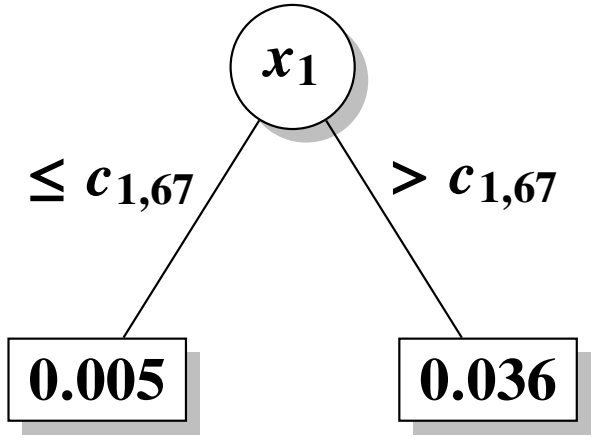


The **BART** R package and trees

Sparapani, Spanbauer and 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 -0.001032108
4    2   0   0  0.004806880
5    3   0   0  0.035709372
6    3  NA  NA            NA
```



Friedman's partial dependence function (FPD) and Marginal Effects of Independent Variables

Suppose that we have a complex regression function, $f(x_S, x_C)$, where x_S is a covariate subset of interest (at a fixed setting) and x_C are the complementary covariates

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

$f_S(x_S)$ is the marginal effect of x_S

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

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

the partial dependence function

where x_{iC} are the training values

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

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

Friedman 2001 *Annals of Statistics*

Friedman's partial dependence function (FPD) and Marginal Effects of Dependent Variables

- Consider our growth chart for height example
- Age and weight obviously co-vary
- t for age, u for sex, v for race/ethnicity and w for weight
  $f_{t,u}(t,u) = \mathbf{E}_{v,w}\left[f(t,u,v,w)|t,u\right]$ assuming Independence
- To do this right, first consider the likely strong relationship between age, sex and weight among children
  $\mathbf{E}\left[w|t,u\right] = \tilde{w} = \tilde{f}(t,u)$
- We can summarize the relationship with a BART model
  $w_i = \tilde{f}(t_i,u_i) + \tilde{\epsilon}_i$ **where** $\tilde{f} \overset{\text{prior}}{\sim}$ **BART**
- A marginal effect more appropriate for dependent variables

$$f_{t,u}(t,u) = \mathbf{E}_v\left[f(t,u,v,\tilde{w})|t,u,\tilde{w} = \mathbf{E}[w|t,u]\right] \quad \text{assuming}$$
$$= \mathbf{E}_v\left[f(t,u,v,\tilde{f}(t,u))|t,u\right] \quad \text{Dependence}$$

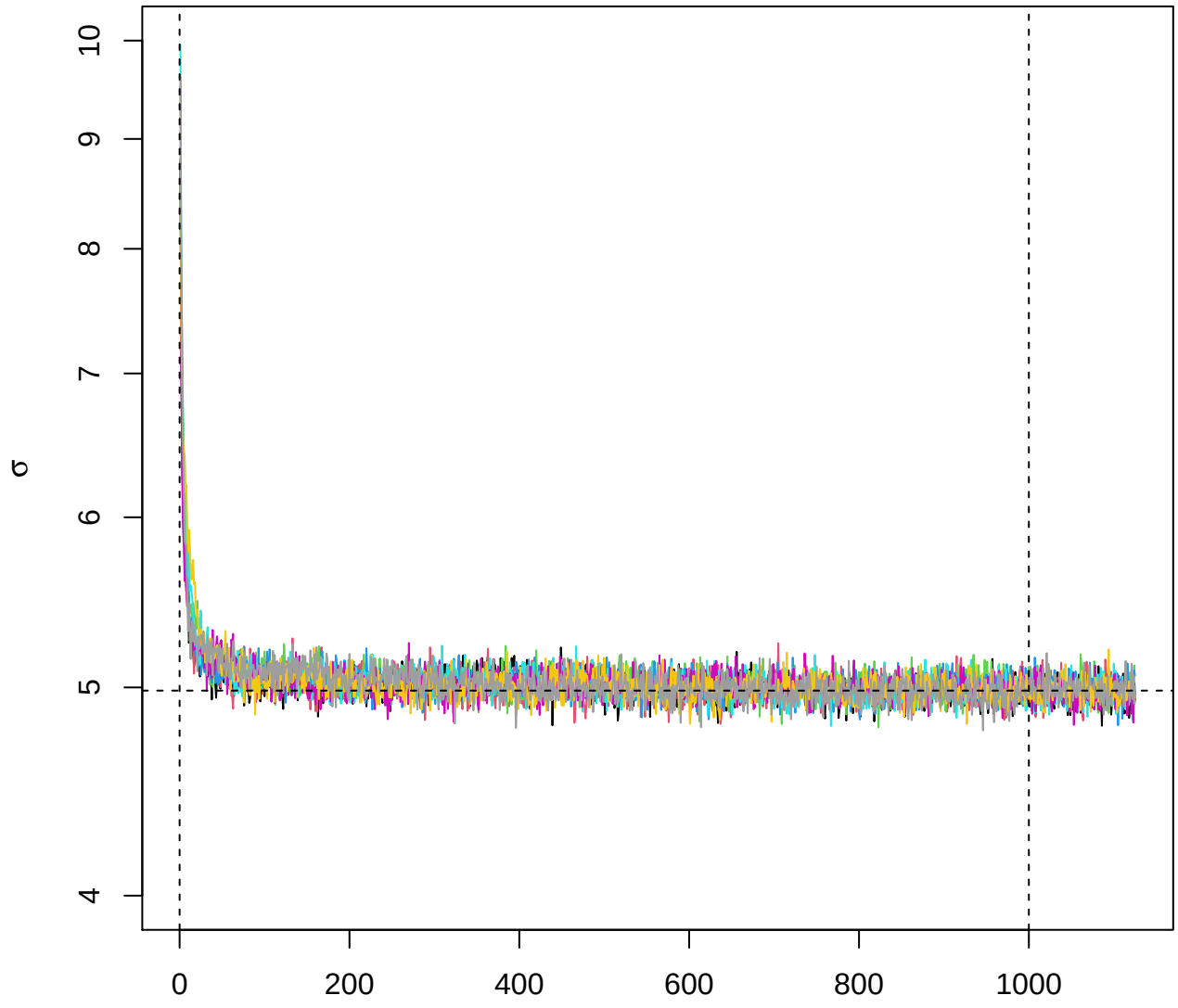
Returning to the real data example

- The 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 `growth1.R,growth2.R,growth3.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*

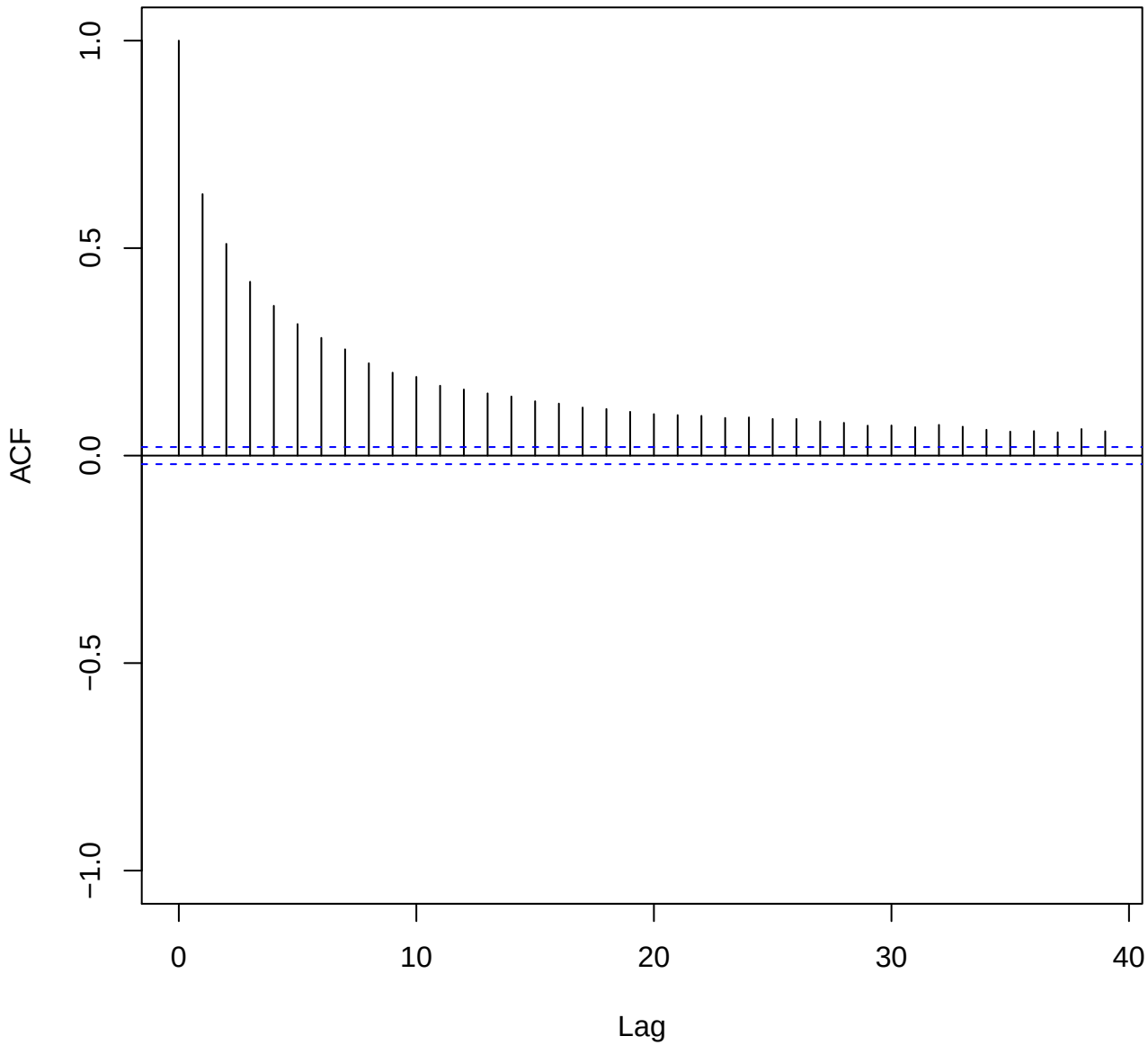
	n	%
Total	3435	
Males	1768	51.5
Females	1667	48.5
White	800	23.3
Black	1035	30.1
Hispanic	1600	46.6

MCMC Convergence `fit$sigma`: **max $\widehat{R}$ = 1.08**

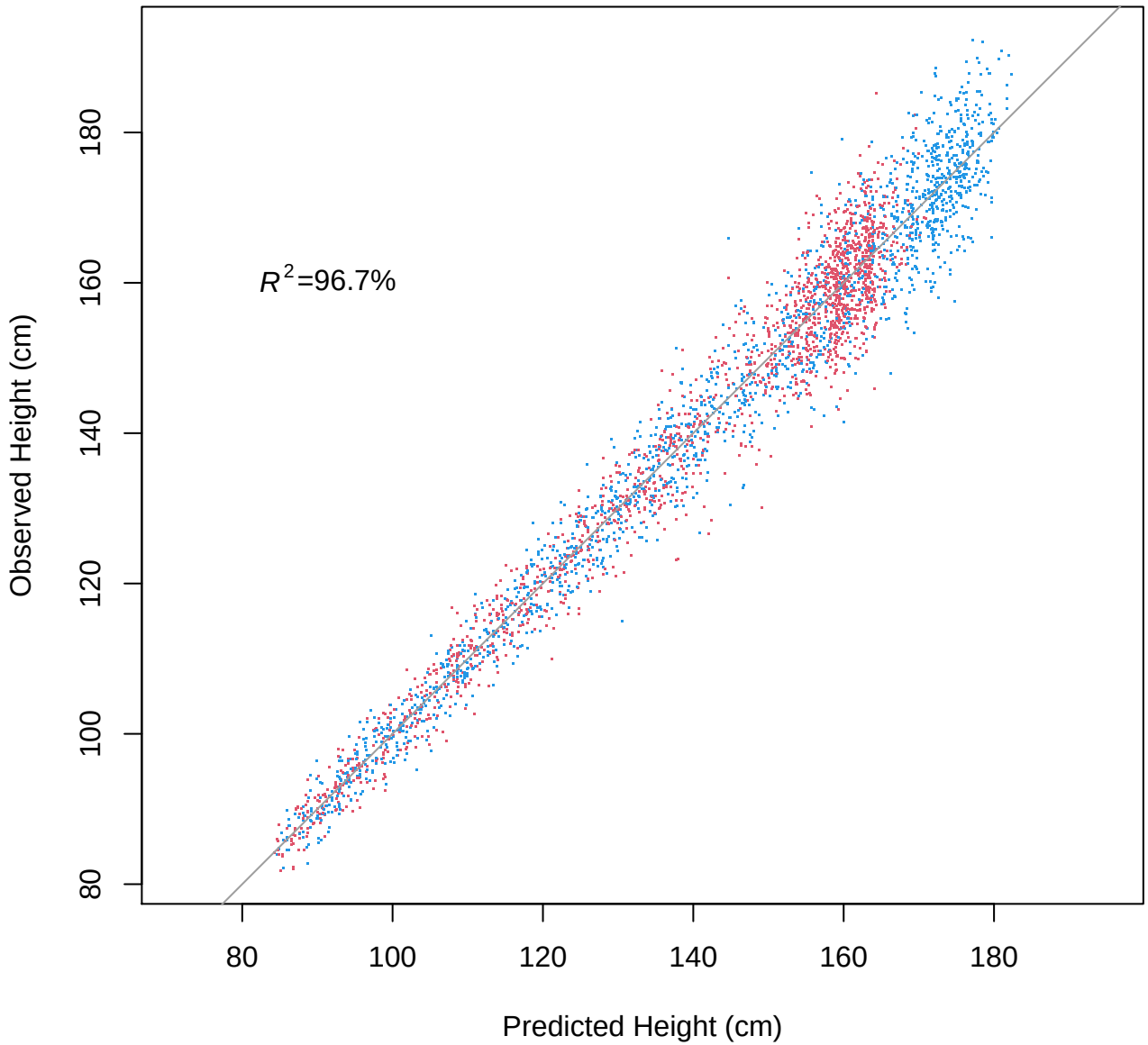
Burn-in 1000, Thinning 10, Chains 8, Posterior 1000



MCMC Convergence `fit$sigma`: Auto-correlation

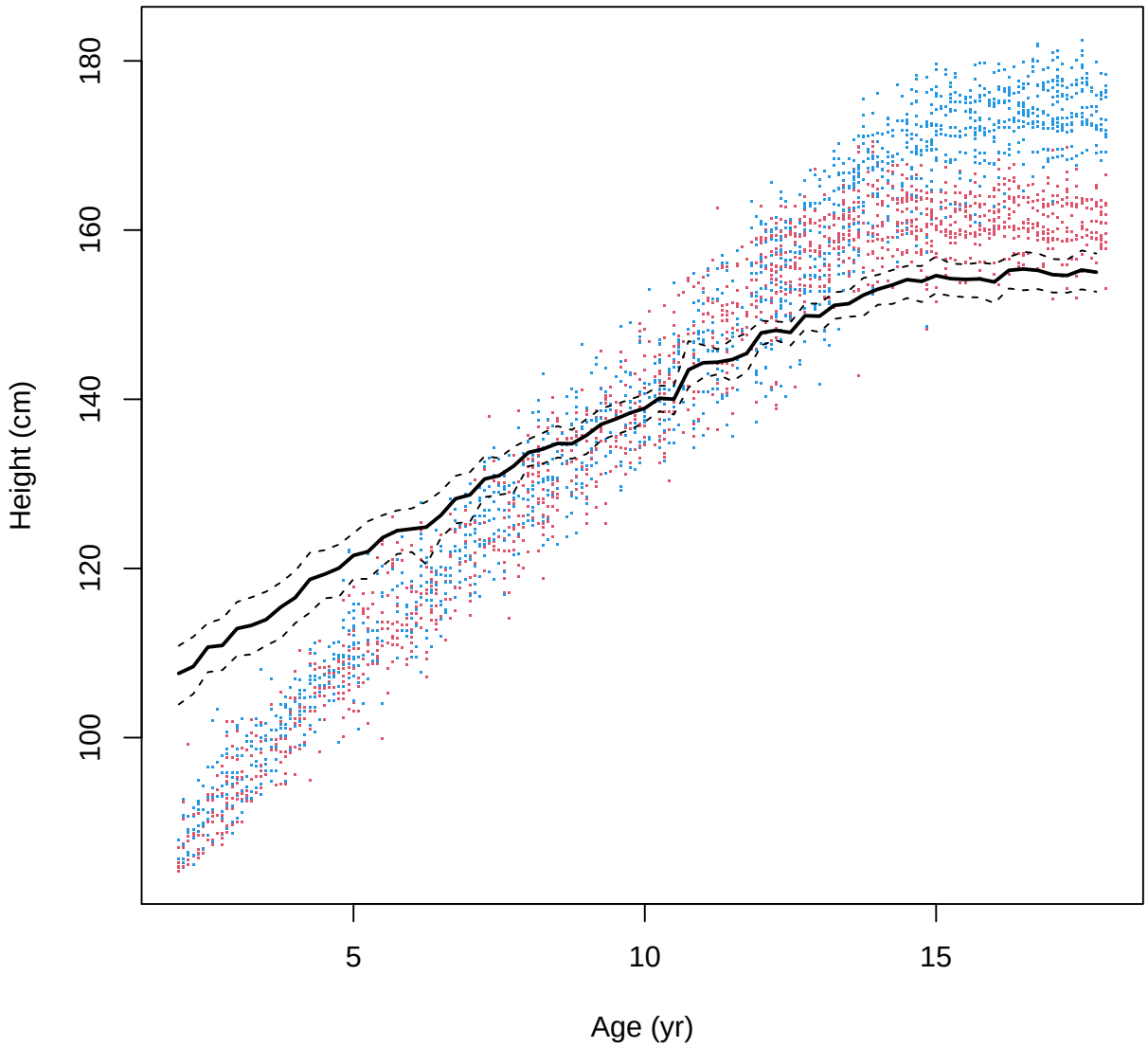


BART fit: M and F

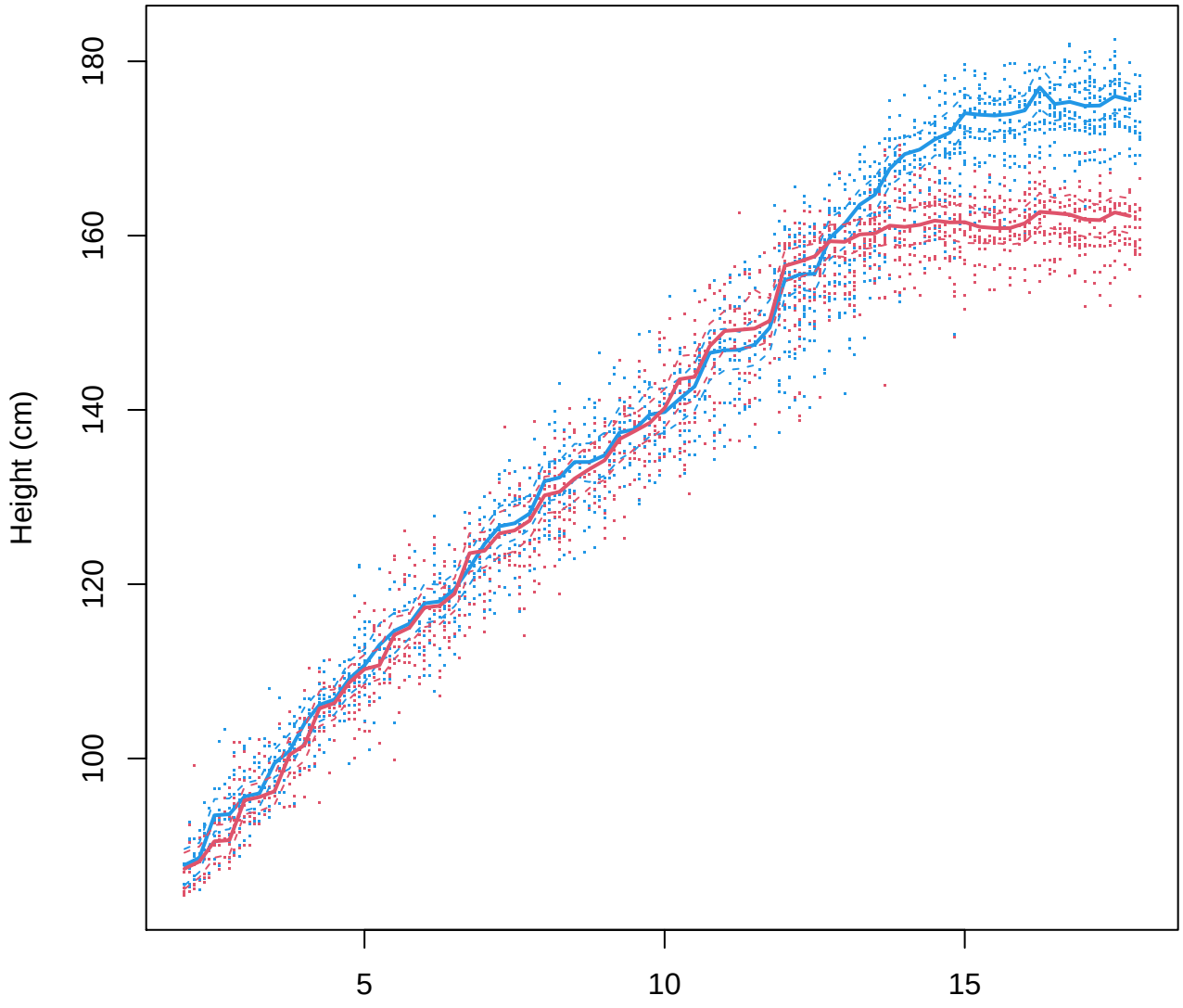


Marginal effect of age assuming weight is independent

$H = 200$, `numcut` $= 100$, BART3 `demo/growth2.R`



Marginal effect of age: BART predictions for M and F assuming weight is dependent, **BART3** `demo/growth3.R`



Heteroskedastic BART (HBART)

Pratola, Chipman, George & McCulloch 2020 *JCGS*

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

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

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

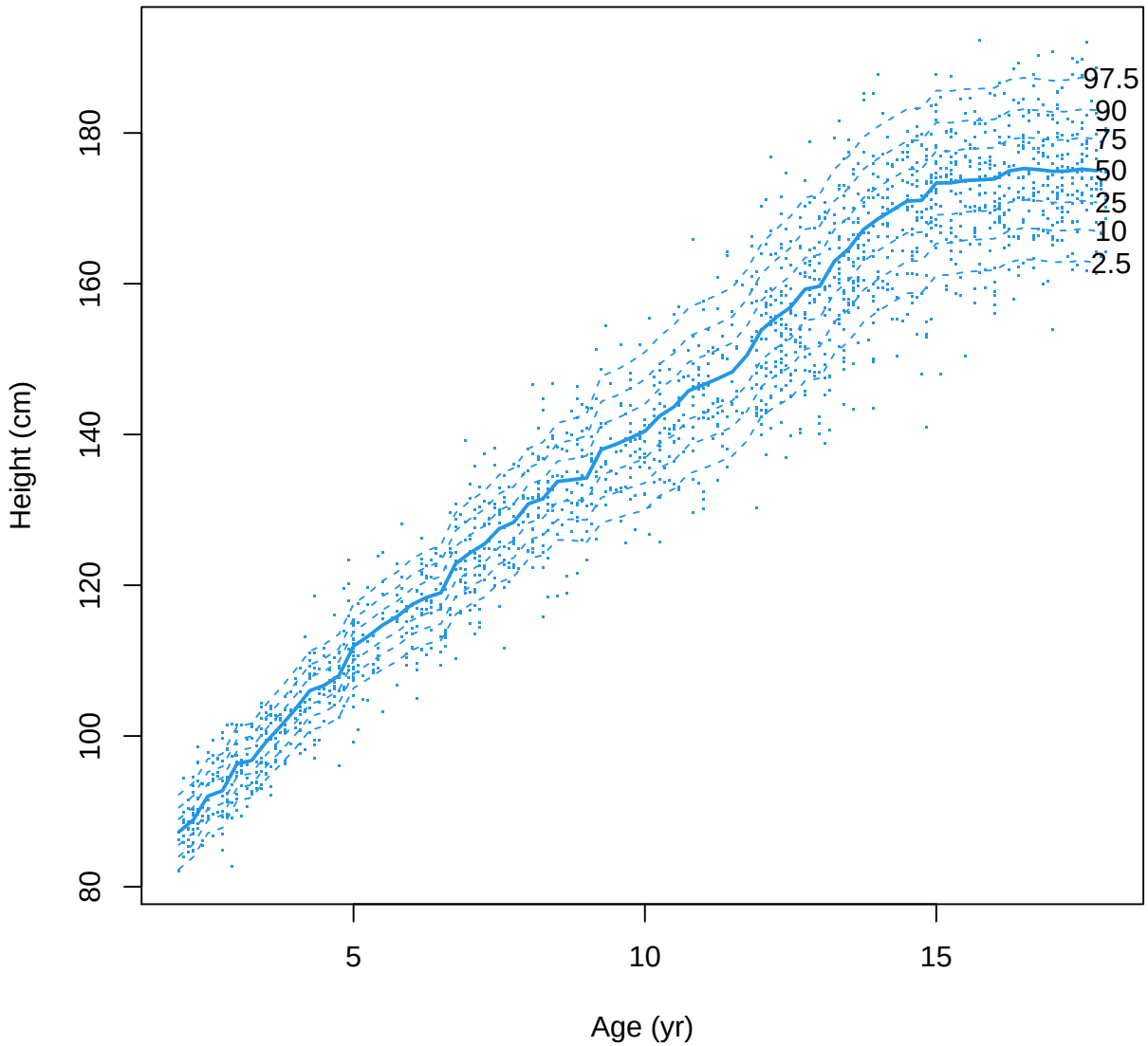
$$s^2(x_i) \equiv \prod_{h=1}^{\widetilde{H}} g(x_i; \widetilde{\mathcal{T}}_h, \widetilde{\mathcal{M}}_h) \qquad \widetilde{H} \approx H/5$$

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

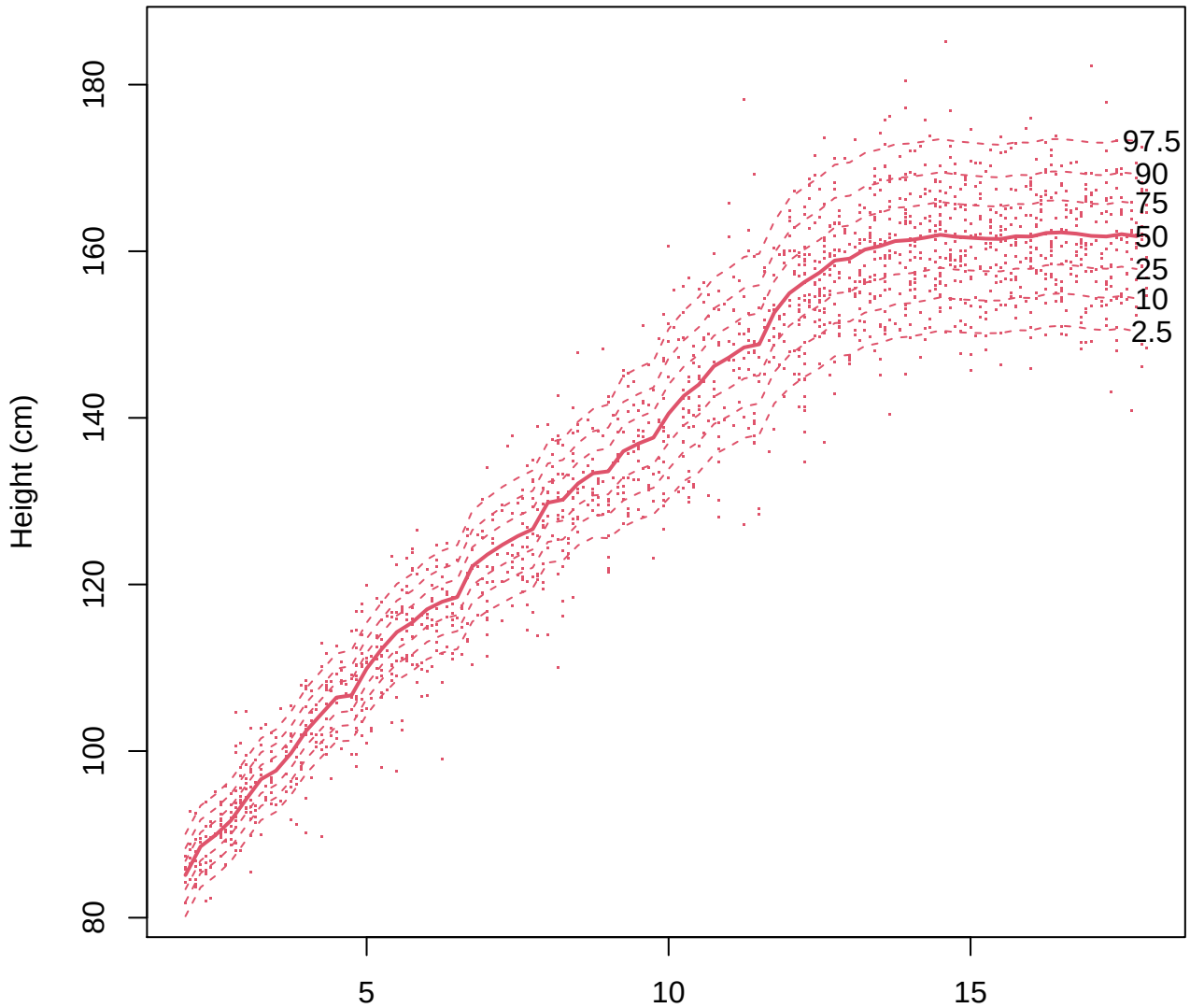
$$\in \widetilde{\mathcal{M}}_h \qquad \nu = 2\left[1 - \left(1 - \frac{2}{\tilde{\nu}}\right)^{1/\widetilde{H}}\right]^{-1}$$

Marginal effect of age: HBART predictions for M

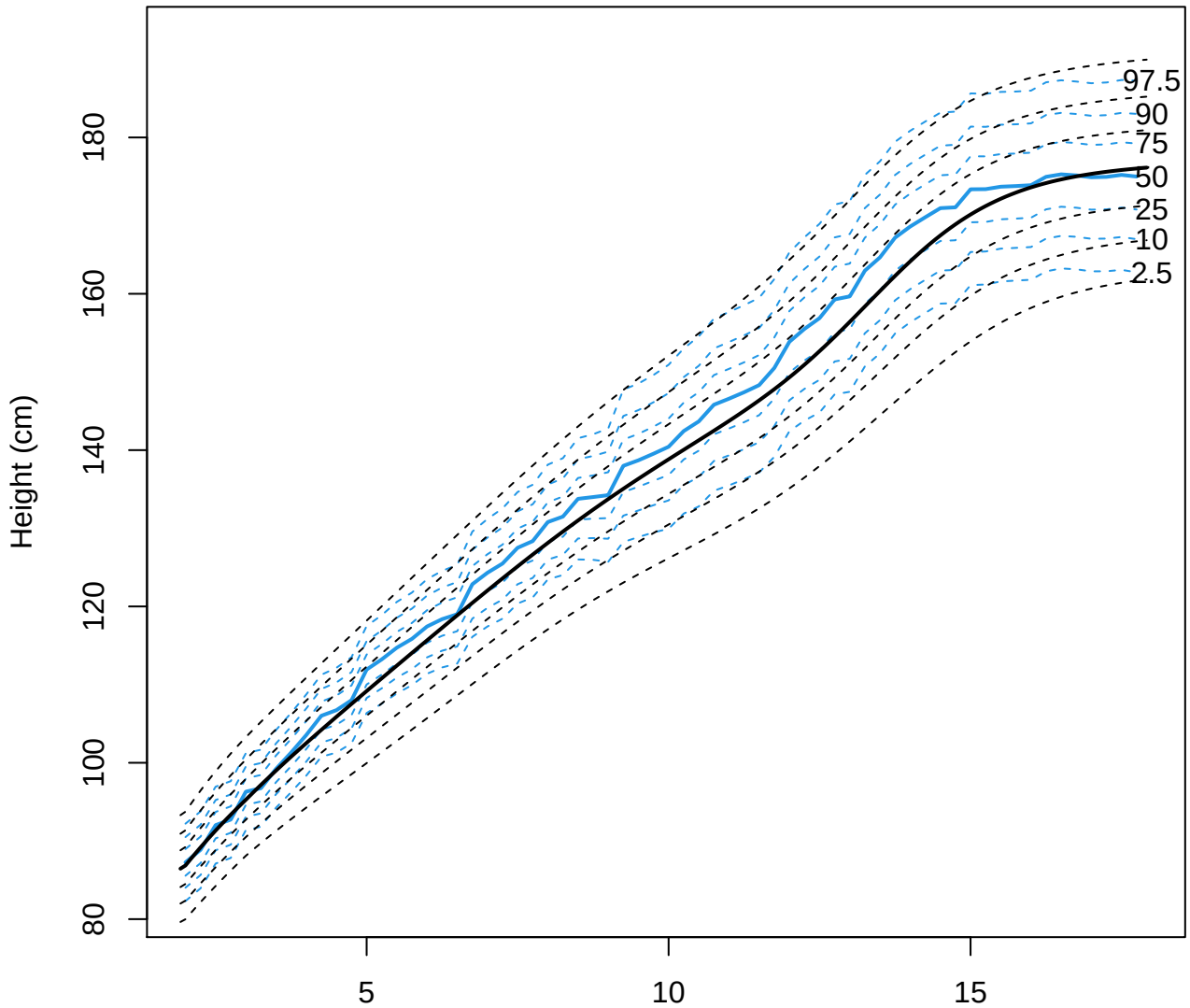
$$H = 300, \widetilde{H} = 60, \texttt{numcut} = 200$$



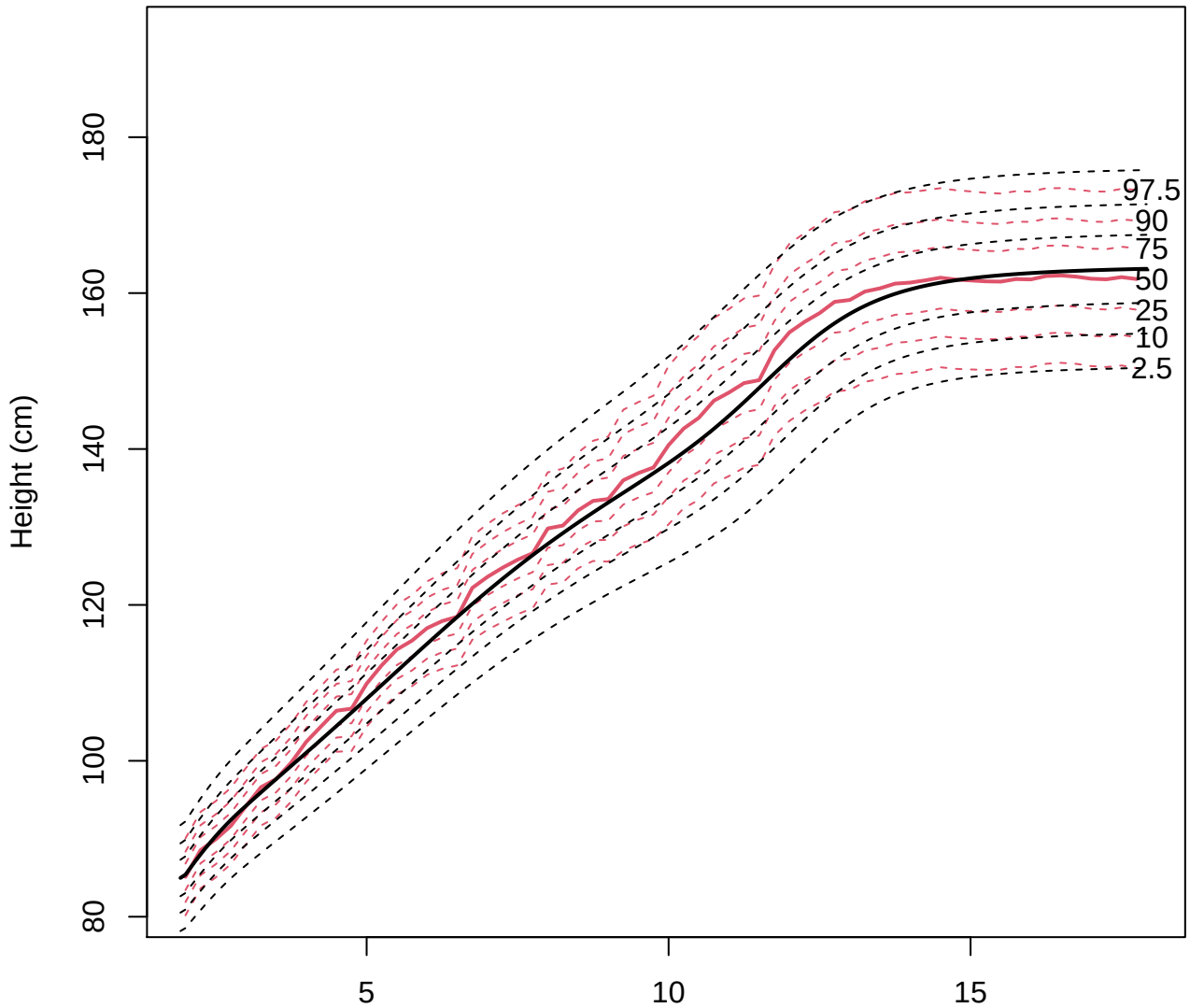
Marginal effect of age: HBART predictions for F assuming weight is dependent, **hbart** `demo/height`



Marginal effect of age: HBART vs. CDC for M assuming weight is dependent, **hbart** `demo/height`



Marginal effect of age: HBART vs. CDC for F assuming weight is dependent, **hbart** `demo/height`



Marginal effects and computational efficiency

- In machine learning, *Shapley values (SHAP)* are another choice for marginal effects (as opposed to FPD)
- However, SHAP are far more computationally intensive than FPD
- Therefore, we do not consider SHAP as a reasonable alternative
- Rather, we want a method faster than FPD that can be tedious
- In fact, we can speed up FPD with *kernel sampling*
  Lundberg and Lee 2017; Janzing, Minorics and Blobaum 2020
- Kernel sampling can also speed up SHAP making it relevant
- The **BART3** package has reliable S3 methods for `FPD` and `FPDK` with kernel sampling: documentation is under construction
- And preliminary support for `SHAP` and `SHAPK`
- Today, I'm going to focus only on `FPD` and `FPDK`

FPDK: FPD with kernel sampling

FPD

$$f_{S_{F_m}}(x_S) \equiv N^{-1} \sum_i f_m(x_S, x_{iC}) \quad x_{iC} \text{ are the training values}$$

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

FPD with kernel sampling

$$f_{S_{K_m}}(x_S) \equiv K^{-1} \sum_k f_m(x_S, x_{kC}) \quad x_{kC} \text{ are random draws of the training}$$

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

FPDK and the kernel sampling empirical variance

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

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

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

FPDK and the kernel sampling empirical variance

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

- 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 $\widehat{s^2}_{S_K(x_S)} = M^{-1}\sum_m s^2_{S_{Km}(x_S)}$
- Therefore, we can empirically estimate the variance like so $\mathbf{V}\left[\hat{f}_{S_F}(x_S)|y\right] \approx \mathbf{V}\left[\hat{f}_{S_K}(x_S)|y\right] - K^{-1}\widehat{s^2}_{S_K(x_S)}$
- So, we generate the posterior for the kernel sampling estimator as $f_{S_{Fm}}(x_S) \approx \hat{f}_{S_K}(x_S) + \left[f_{S_{Km}}(x_S) - \hat{f}_{S_K}(x_S)\right]\sqrt{\frac{\mathbf{V}\left[\hat{f}_{S_F}(x_S)|y\right]}{\mathbf{V}\left[\hat{f}_{S_K}(x_S)|y\right]}}$