Introduction to BART with binary outcomes

Rodney Sparapani
Medical College of Wisconsin

September 30: BART Bootcamp
Biostatistics in the Modern Computing Era
Medical College of Wisconsin, Milwaukee campus

Funding for this research was provided, in part, by the Advancing Healthier Wisconsin Research and Education Program under awards 9520277 and 9520364.

Outline

- Motivation: chronic spine pain and obesity
- Probit regression with BART
- Logistic regression with BART
- Geweke convergence diagnostics for binary BART

Motivation: chronic spine pain and obesity

- We believe that obesity is a risk factor for chronic lower back/buttock pain
- Conversely, obesity is not a risk factor for chronic neck pain
- Data available from the National Health and Nutrition Examination Survey (NHANES) 2009-2010 Arthritis Questionnaire
- 5106 subjects were surveyed
- Demographics: age and gender
- Anthropometrics available: weight (kg), height (cm), body mass index (kg/m^2), waist circumference (cm)
- Sampling weights to estimate for the US as a whole

Probit BART for binary outcomes

Probit regression with latent variables: Albert & Chib 1993 *JASA*

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

$$p_i|f = \Phi(\mu + f(x_i)) \text{ where } f \overset{\text{prior}}{\sim} \text{BART}$$

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

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

$$[y|f] = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i} \qquad \text{Likelihood}$$

Continuous BART with unit variance, $\sigma^2 = 1$, and z_i are the data

Friedman's partial dependence function

Friedman 2001 *AnnStat*

$$f(x) = f(x_S, x_C) \qquad \text{BART function where } x = [x_S, x_C]$$

$$f(x_S) = \mathbf{E}_{x_C}\left[f(x_S, x_C)|x_S\right]$$

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

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

$$\hat{f}(x_S) \equiv M^{-1}\sum_m f_m(x_S)$$

`pbart` and `mc.pbart` input and output

```
post <- pbart(x.train, y.train, ...,
        ndpost=M, keepevery=1) or
post <- mc.pbart(x.train, y.train, ...,
        ndpost=M, keepevery=1, mc.cores=2, seed=99)
```

Input matrices: `x.train` and, optionally, `x.test`: x_i

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

Output object, `post`, of type `pbart` which is essentially a list

Matrices: `post$yhat.train` and `post$yhat.test`: $\hat{y}_{im} = f_m(x_i)$

$$\begin{bmatrix} \hat{y}_{11} & \cdots & \hat{y}_{N1} \\ \vdots & \vdots & \vdots \\ \hat{y}_{1M} & \cdots & \hat{y}_{NM} \end{bmatrix}$$

`predict.pbart` input and output

```
pred <- predict(post, x.test, mc.cores=1, ...)
```

Input matrices: `x.test`: x_i

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_Q \end{bmatrix}$$

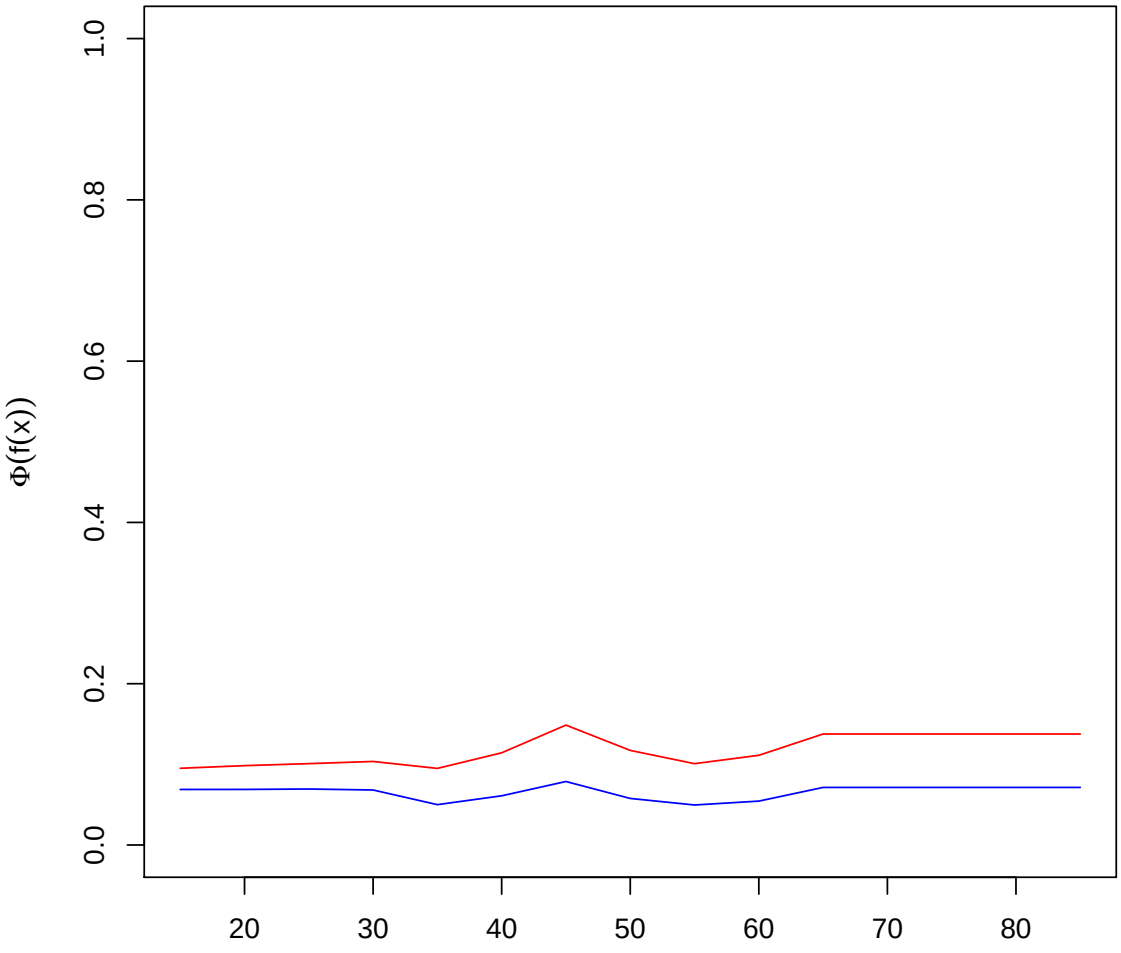
Output matrix `pred`: $\hat{y}_{im} = f_m(x_i)$

$$\begin{bmatrix} \hat{y}_{11} & \cdots & \hat{y}_{Q1} \\ \vdots & \vdots & \vdots \\ \hat{y}_{1M} & \cdots & \hat{y}_{QM} \end{bmatrix}$$

Live demo: chronic spine pain and obesity

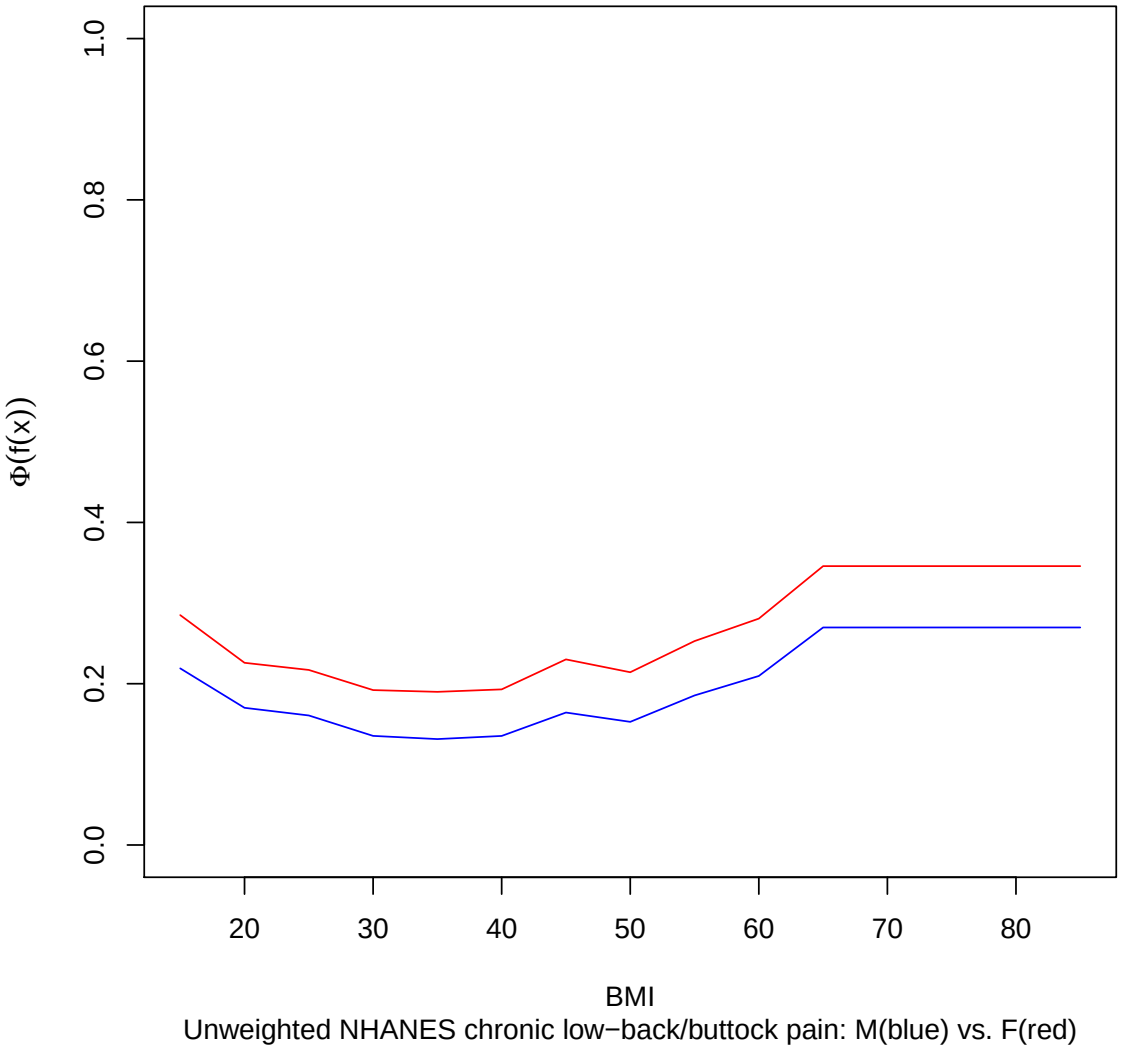
- `system.file('demo/nhanes.pbart1.R', package='BART')`
- `system.file('demo/nhanes.pbart2.R', package='BART')`

Friedman's partial dependence function: Probability of chronic neck pain vs. BMI



Unweighted NHANES chronic neck pain: M(blue) vs. F(red)

Friedman's partial dependence function:
Probability of chronic lower back/buttock pain vs. BMI



Logistic BART for binary outcomes

Logistic regression: Holmes & Held 1993 *Bayesian Analysis*
Devroye 1986 *Non-uniform random variate generation*

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

$$p_i|f = \Phi(f(x_i)) \text{ where } f \overset{\text{prior}}{\sim} \text{BART}$$

$$z_i|y_i,f \sim \mathbf{N}\big(f(x_i),\ \sigma_i^2\big) \begin{cases} \mathbf{I}(-\infty,0) & \text{if } y_i = 0 \\ \mathbf{I}(0,\infty) & \text{if } y_i = 1 \end{cases}$$

$$\sigma_i^2 = 4\psi_i^2 \text{ where } \psi_i \sim \text{Kolmogorov-Smirnov (see Devroye)}$$

Continuous BART with heteroskedastic variance and z_i is the data

Geweke convergence diagnostics for binary BART

Hastings 1970 *Biometrika*, Silverman 1986 *Chapman and Hall*

$$\hat{\theta}_M = M^{-1} \sum_{m=1}^{M} \theta_m$$ Bayesian estimator

$$\sigma^2_{\hat{\theta}} = \lim_{M \to \infty} \mathbf{V}\left[\hat{\theta}_M\right]$$ Asymptotic variance

Suppose θ_m is an **ARMA** (p, q)

$$\gamma(w) = (2\pi)^{-1} \sum_{m=-\infty}^{\infty} \mathbf{V}\left[\theta_0, \theta_m\right] \mathbf{e}^{imw}$$ Spectral density

$$\hat{\sigma}^2_{\hat{\theta}} = \hat{\gamma}^2(0)$$ Variance estimator

Geweke convergence diagnostics for binary BART

Geweke 1992 *Bayesian Statistics*

- Divide your chain into two segments: A and B
- $m \in A = \{1, \ldots, M_A\}$ where $M_A = aM$
- $m \in B = \{M - M_B + 1, \ldots, M\}$ where $M_B = bM$
- $a + b < 1$, Geweke suggests $a = 0.1$ and $b = 0.5$

$$\hat{\theta}_A = M_A^{-1} \sum_{m \in A} \theta_m \qquad \hat{\theta}_B = M_B^{-1} \sum_{m \in B} \theta_m$$

$$\hat{\sigma}^2_{\hat{\theta}_A} = \hat{\gamma}^2_{m \in A}(0) \qquad \hat{\sigma}^2_{\hat{\theta}_B} = \hat{\gamma}^2_{m \in B}(0)$$

$$z = \frac{\sqrt{M}(\hat{\theta}_A - \hat{\theta}_B)}{\sqrt{a^{-1}\hat{\sigma}^2_{\hat{\theta}_A} + b^{-1}\hat{\sigma}^2_{\hat{\theta}_B}}} \qquad \sim \mathrm{N}(0,\ 1)$$

Geweke convergence diagnostics for binary BART

- We have a z_i corresponding to each $\theta_i = h(f(x_i))$
- In the **BART** R package, we created the `gewekediag` function which was adapted from the **coda** R package
  Plummer, Best et al. 2006

```
system.file('demo/geweke.pbart2.R', package='BART') &
system.file('demo/geweke.pbart3.R', package='BART')
```

Geweke convergence diagnostics for binary BART: simulated data scenario

```
system.file('demo/geweke.pbart2.R', package='BART') &
system.file('demo/geweke.pbart3.R', package='BART')
```

$$N = 100, 1000, 10000$$

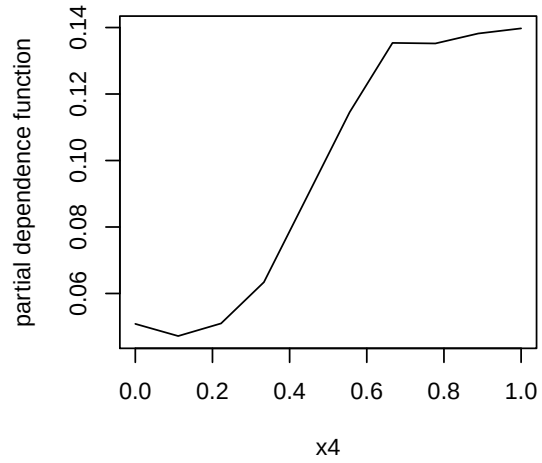
$$K = 50$$

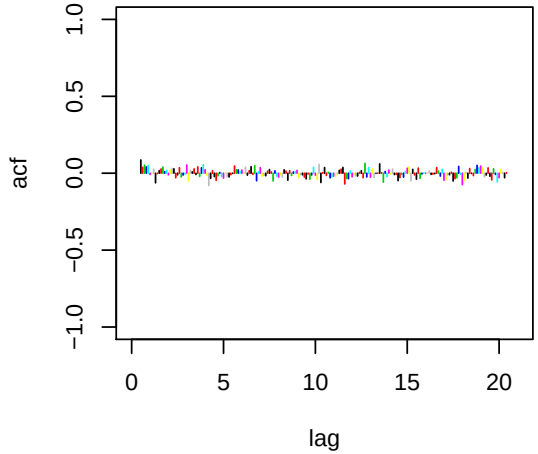
$$f(x_i) = -1.5 + \sin(\pi x_{1i} x_{2i}) + 2(x_{3i} - 0.5)^2 + x_4 + 0.5x_5$$

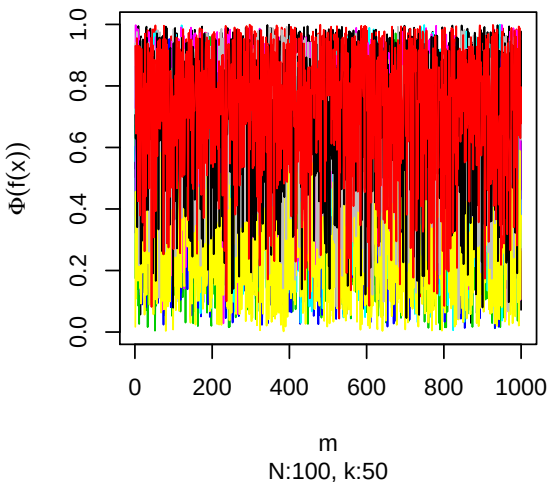
$$z_i \sim \mathrm{N}(f(x_i),\ 1)$$

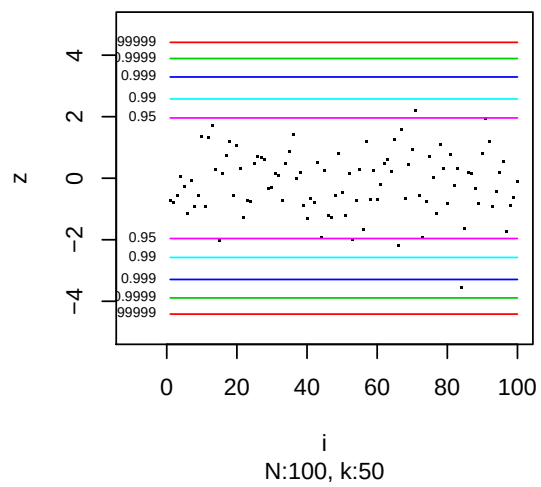
$$y_i = \mathrm{I}(z_i > 0)$$

Geweke convergence diagnostics for binary BART: $N = 100$

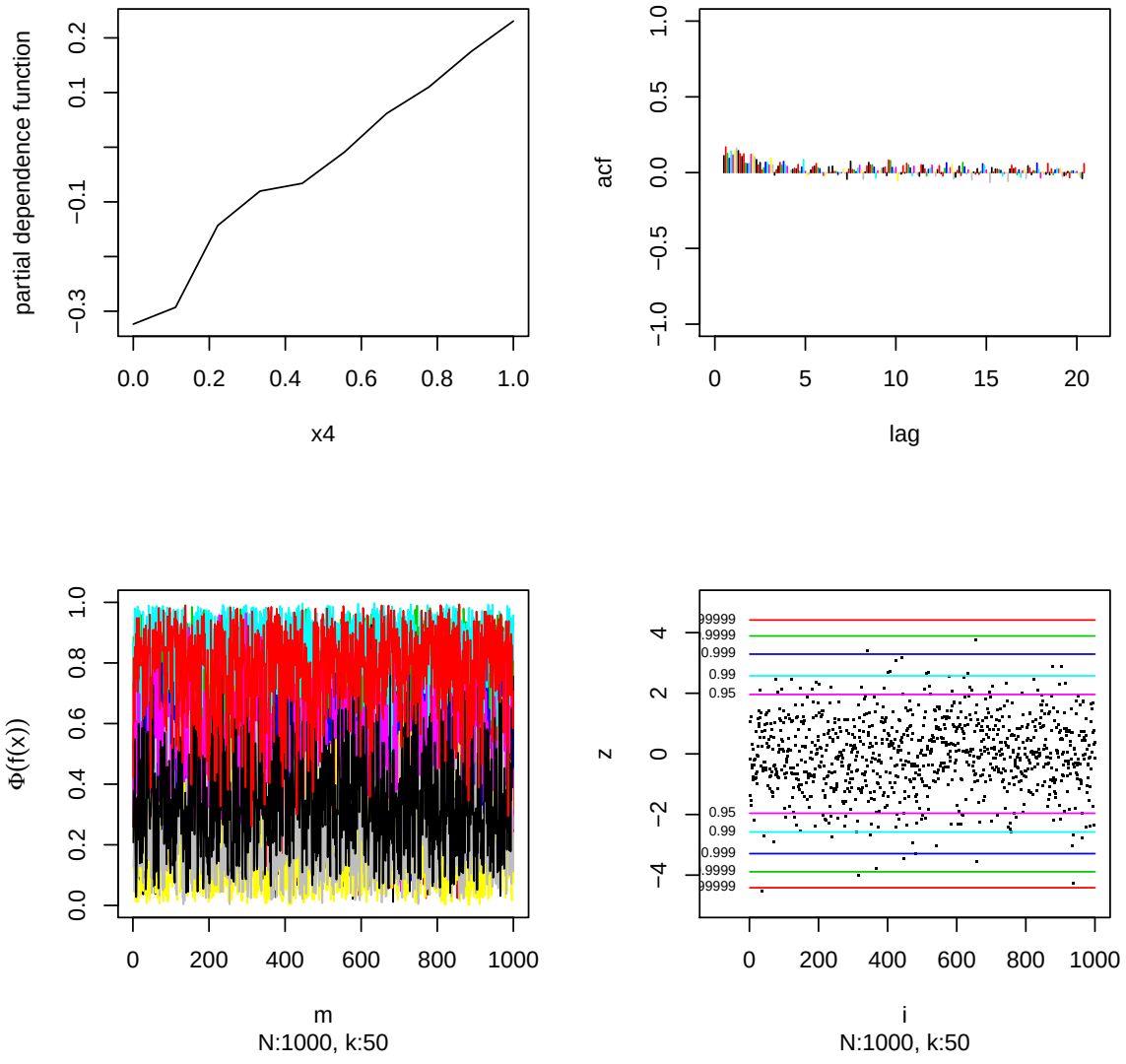




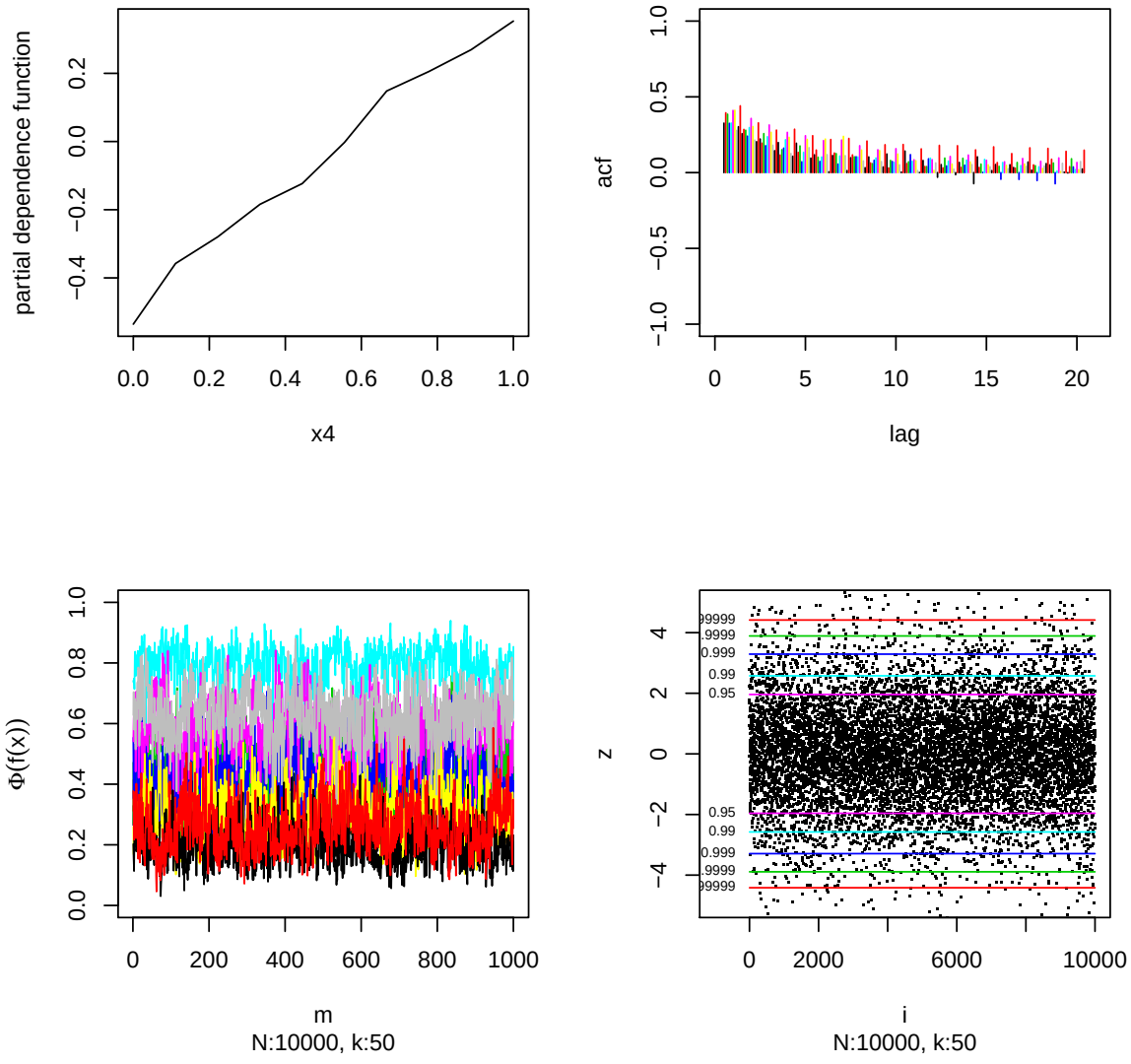




Geweke convergence diagnostics for binary BART: $N = 1000$



Geweke convergence diagnostics for binary BART: $N = 10000$

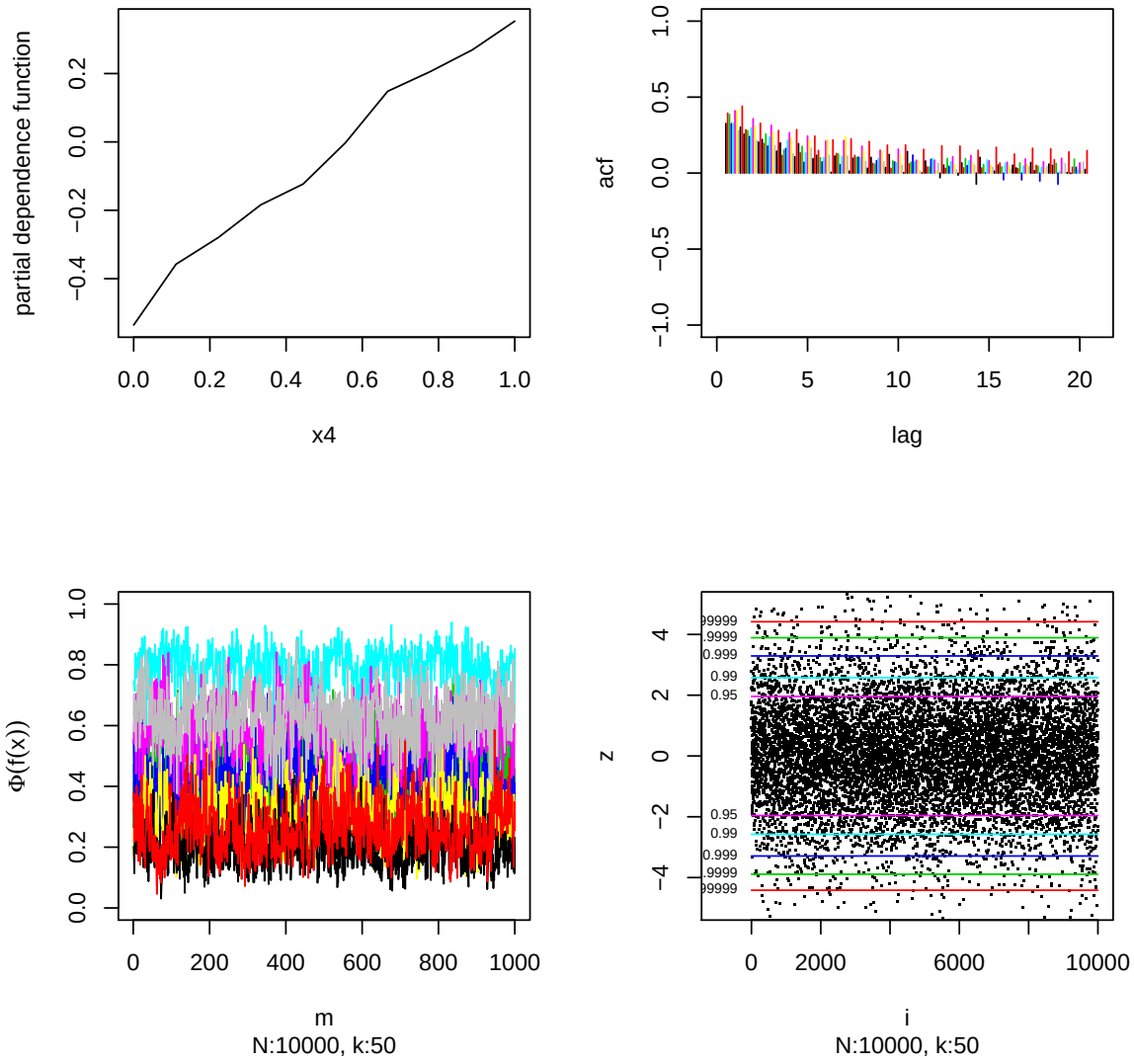


Geweke convergence diagnostics for binary BART: convergence countermeasures for $N = 10000$

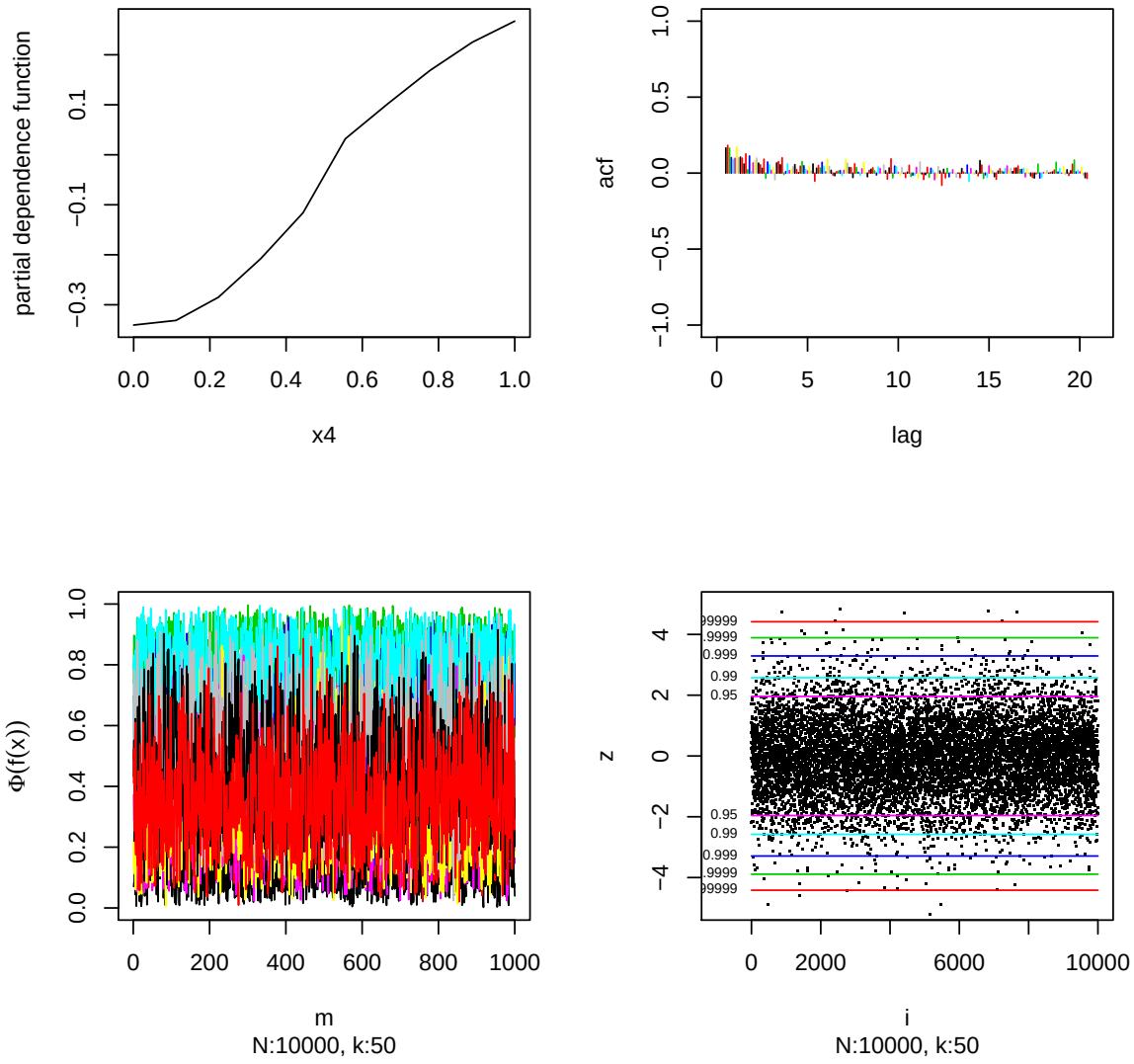
```
system.file('demo/geweke.pbart3.R', package='BART')
```

- Validation set is a 10% random sample of the training set
- Randomly partition the training set into 10 sets of $N = 1000$
- Analyze the 10 training sets independently with BART
- Combine/stack the 10 validation set MCMC chains
- Perform convergence diagnostics and inference on the combined validation chains

Geweke convergence diagnostics for binary BART: $N = 10000$ without countermeasures



Geweke convergence diagnostics for binary BART: $N = 10000$ with countermeasures



Geweke convergence diagnostics for binary BART: $N = 1000$ without countermeasures