Chipman H, George E, Hahn R, McCulloch R, Pratola M, Sparapani, R. “Computational approaches to Bayesian Additive Regression Trees”. Chapter of Computational Statistics in Data Science. Piegorsch W, Levine R, Zhang HH, Lee TCM (eds.) Wiley, 2022.

# 5 Example: Boston housing values and air pollution

Here, we demonstrate BART with the classic Boston housing example [12]. This data is based on the 1970 US Census where each observation represents a Census tract in the Boston Standard Metropolitan Statistical Area. For each tract, there was a localized air pollution estimate, the concentration of nitrogen oxides, `nox`, based on a meteorological model that was calibrated to monitoring data. Restricted to tracts with owner-occupied homes, there are $N = 506$ observations. We’ll predict the median value of owner-occupied homes (in thousands of dollars), `mdev`, by thirteen covariates including `nox` which is our primary interest.

However, BART does not directly provide a summary of the effect of a single covariate, or a subset of covariates, on the outcome. Friedman’s partial dependence function [9] can be employed with BART to summarize the marginal effect due to a subset of the covariates, $\boldsymbol{x}_S$, by aggregating over the complement covariates, $\boldsymbol{x}_C$, i.e., $\boldsymbol{x} = [\boldsymbol{x}_S, \boldsymbol{x}_C]$. The marginal dependence function is defined by fixing $\boldsymbol{x}_S$ while aggregating over the observed settings of the complement covariates in the data set: $f(\boldsymbol{x}_S) = N^{-1} \sum_{i=1}^{N} f(\boldsymbol{x}_S, \boldsymbol{x}_{iC})$. For example, suppose that we want to summarize `mdev` by `nox` while aggregating over the other twelve covariates in the Boston housing data. In Figure 2, we demonstrate the marginal estimate and its 95% credible interval: notice that BART has discerned a complex non-linear relationship between `mdev` and `nox` from the data. N.B. this example including data and source code can be found in the `BART` `R` package [24] as the `nox.R` demonstration program.

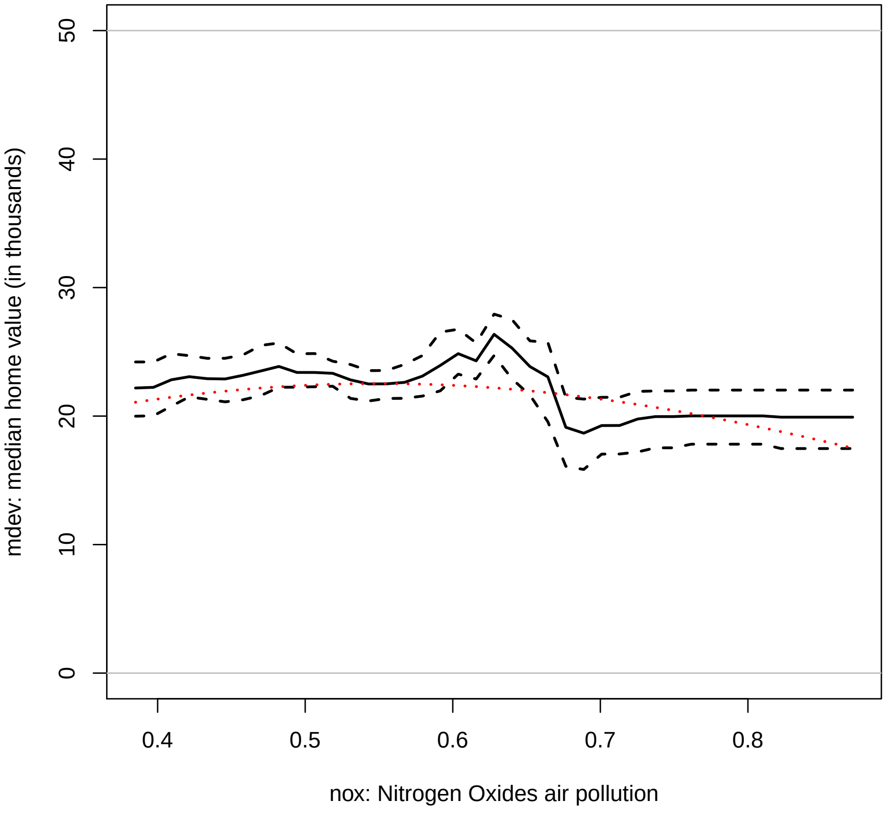

Figure 2: The Boston housing data was compiled from the 1970 US Census where each observation represents a Census tract in Boston with owner-occupied homes. For each tract, we have the median value of owner-occupied homes (in thousands of dollars), `mdev`, and thirteen other covariates including a localized air pollution estimate, the concentration of nitrogen oxides `nox`, which is our primary interest. We summarize the marginal effect of `nox` on `mdev` while aggregating over the other covariates with Friedman's partial dependence function. The marginal estimate and its 95% credible interval are shown. The red line with short dashes comes from the linear regression model of [12] where a quadratic effect of `nox` with respect to the logarithm of `mdev` is assumed.