Residuals and Influence in Bayesian Ensemble Models

Rob McCulloch and Matt Pratola

Arizona State, Ohio State

1. Residuals and Influence in Data Science
2. Cooks' Distance for A Bayesian Single Tree Model
3. Single Tree Model: A Simple Simulated Example in 1D
4. Cooks' Distance for A Bayesian Ensemble
5. Ensemble of Trees: A Simple Simulated Example in 1D
6. Cars Data: A Higher-Dimensional Real Data Example
7. Conclusion

1. Residuals and Influence in Data Science

Applied Linear Regression is still a fundamental part of our toolkit.

Any good Applied Linear Regression course or book emphasizes **diagnostics**:

residuals and influence

Most applied Machine Learning Books do not!!!

While the emphasis on out-of-sample prediction in Machine Learning is a great thing, it does not seem like the motivation for doing diagnostics we learn in applied linear regression is any less relevant for the larger set of predictive tools currently employed.

In this talk we take a class of tree based models we have been working on and

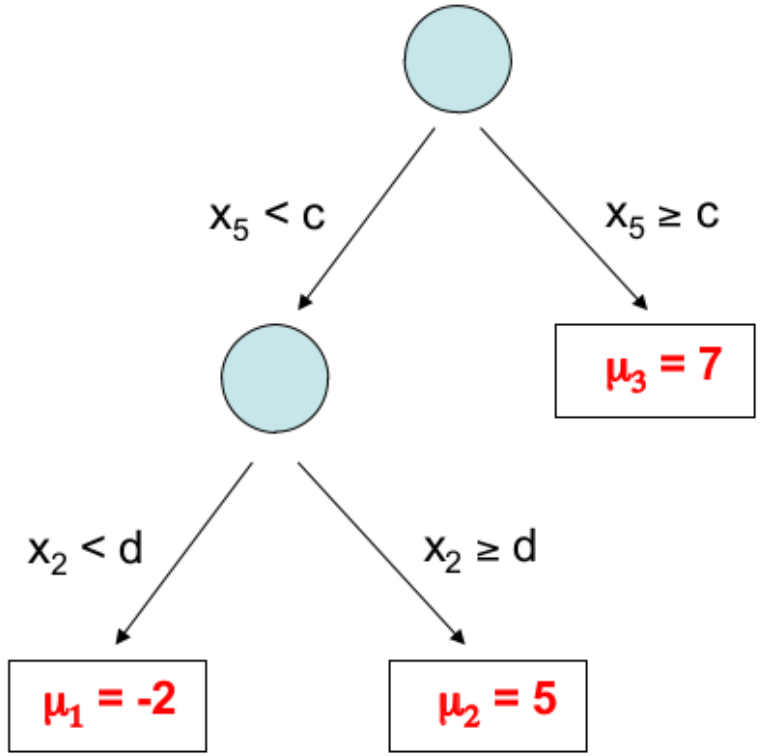
- Propose a simple measure of influence based on Cook's distance.
- Investigate the role of residuals and influence in our models.

A single tree model:

Let T denote the tree structure including the decision rules.

Let $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ denote the set of bottom node μ's.

Let $f(x; T, M)$ be a regression tree function that assigns a μ value to x.



A single tree model:

$$y = f(x; T, M) + \epsilon.$$

The BART model:

$$Y = f(x) + \sigma Z$$

and

$f(x)$ is represented as the sum of *many* single tree models:

$$f(x) = \sum_{i=1}^{m} f(x \mid T_i, M_i)$$

where each (T_i, M_i) represents a single tree model.

m is hundreds , thousands.

2. Cooks' Distance for A Bayesian Single Tree Model

In linear regression, we look at outliers and compute Cook's distance.

Cook's distance helps us identify *influential* outliers, in that removing them could alter the fit enough to make a practical difference.

The formula for Cook's distance in linear regression is a miracle. It is simple, interpretable, and tell us what we want to know.

We can assess the effect on the fit of removing an observation without having to do it.

In some modern approaches, fitting the model to data involves a complex algorithm so that a simple result like Cook's distance is not available.

Example:

Fitting a deep neural net using stochastic gradient descent.

Example:

Fitting a regression tree, or an ensemble of regression trees using Markov Chain Monte Carlo.

We are working on the second example, and that is what we will focus on today.

We could use numerical approximations to approximate the refit after dropping an observation:

Neural Net

Take a few gradient steps away from the solution obtained from the full data using only a subset of the observations.

MCMC for Trees

Reweight the MCMC draws using the likelihood from a subset of observations.

While well worth pursuing, these approaches entail computational issues.

We want a simple approach.

We identify a piece of the model that looks like a linear model and then apply Cooks' distance !!

Our Bayesian MCMC approach gives us draws in the space of decision trees.

Given a tree, we can write our model as

$$y = \sum_{j=1}^{B} I_j \, \mu_j + \epsilon$$

where

- I_j is 1 if the i^{th} observation is in bottom node j, and 0 else.
- μ_j is the mean level assigned to bottom node j.
- ϵ is vector of iid $N(0, \sigma^2)$ errors.

Conditional on the tree, we can write our model as a multiple regression where our regressors are dummy variables indicating which bottom node an observation belongs to.

$$y = \sum_{j=1}^{B} I_j \, \mu_j + \epsilon$$

Our approach is to simply compute Cook's Distance based on the multiple regression for each MCMC draw.

At each draw, the tree can change so that the dummies change.
At each draw, we get a new σ value.

Let $n_{[i]}$ = the number of observations in the bottom node corresponding to the i^{th} observation.

Let B denote the number of bottom nodes.

Then our version of Cook's distance for the i^{th} observation is:

$$D_i = \frac{1}{B}\frac{e_i^2}{\sigma^2}\frac{n_{[i]}}{(1-n_{[i]})^2}.$$

If you just apply a simple Cook's distance formula to our (no intercept) dummy regression you get the above except that:

- We have replaced the least squares $\hat{\sigma}$ with the current MCMC draw of σ.
- e_i is based on the current MCMC draw of the tree (and corresponding indicator regressors) and $\{\mu_j\}$.

Just looking at the posterior distribution of $\frac{e_i}{\sigma}$ would be quite reasonable (as suggested by Zellner).

We borrow a slight adjustment from Cook's distance to account for the number of observations in a bottom node.

Clearly our approach is ad-hoc.
In particular, we ignore the effect of influential observations on the MCMC traversal of tree space.

But it is simple !! (and somewhat interpretable).

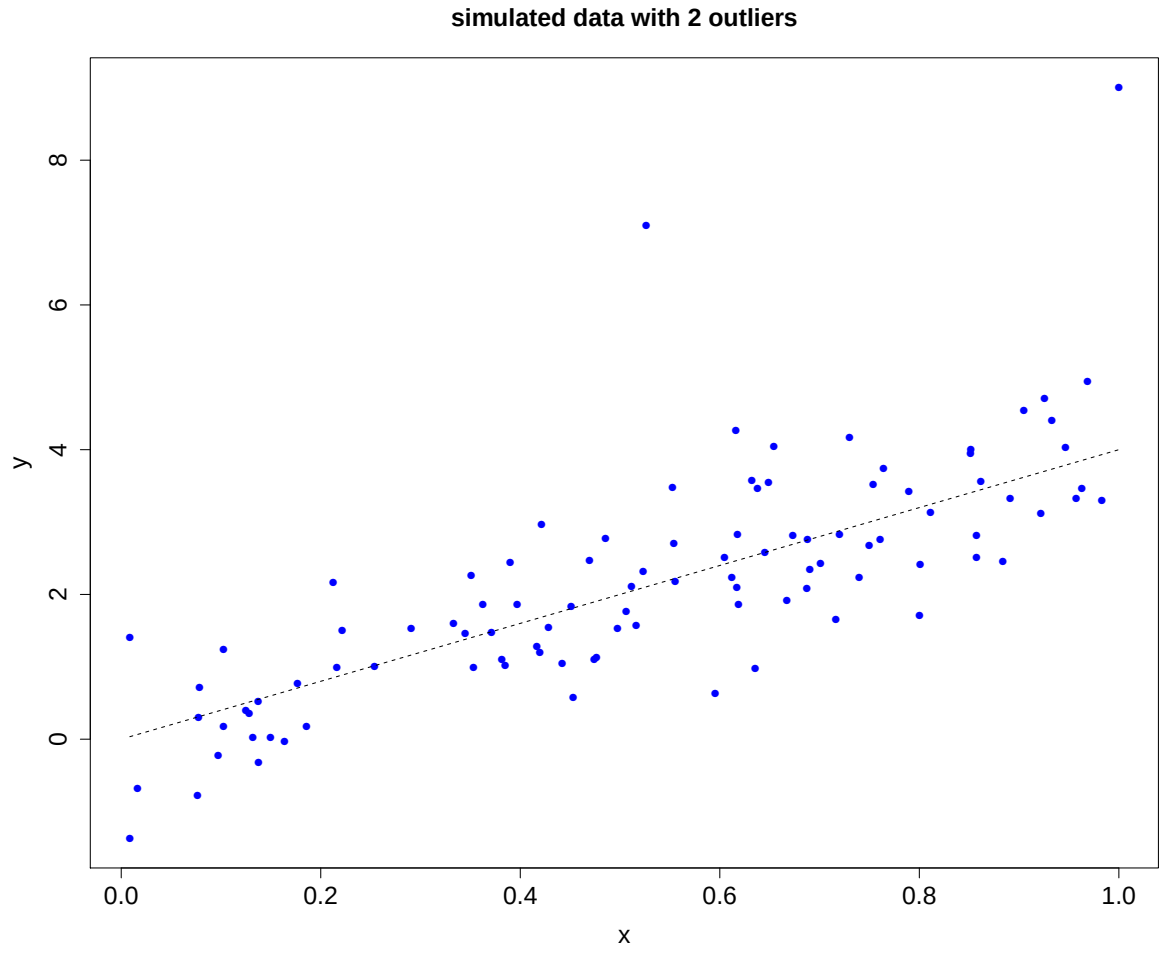
3. Single Tree Model:A Simple Simulated Example in 1D

We simulate a simple regression problem with just one x.

The true relationship is linear.

There are two outliers, one in the middle of the data, and one at the edge.

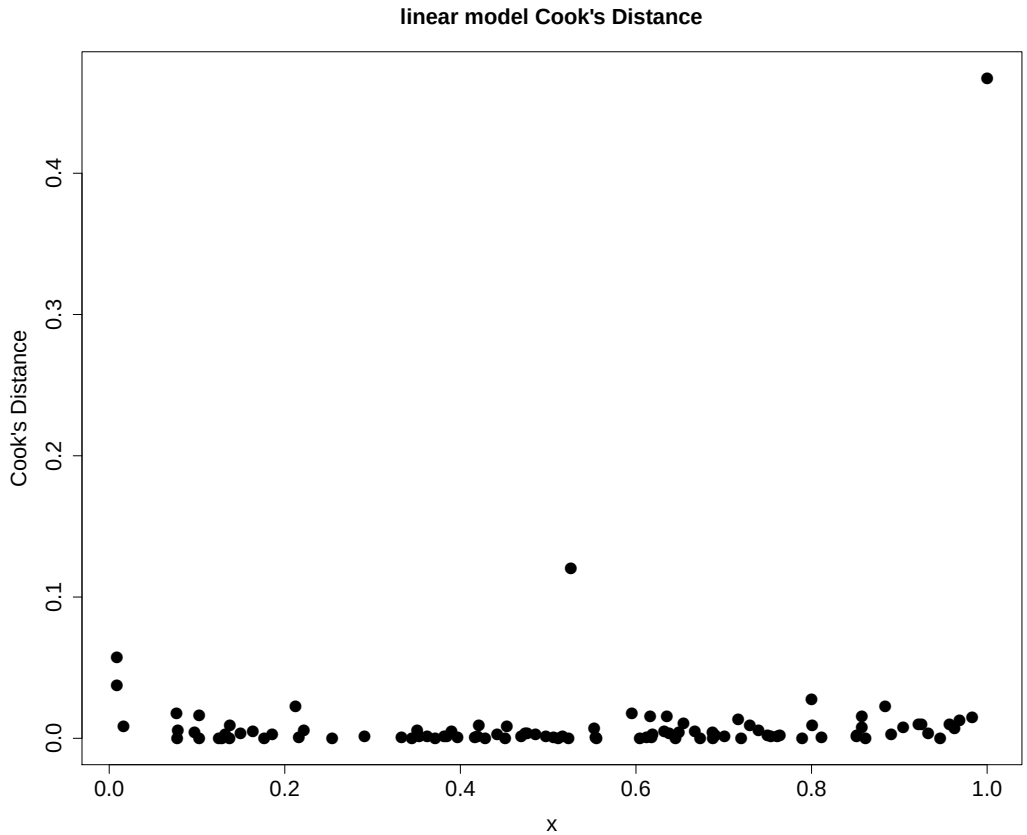
We examine the influence of the outliers on the fit of our tree based models and compare it to the linear case.



Two outliers.

Does the outlier at the edge of the data effect the inference more than the outlier in the middle of the data??

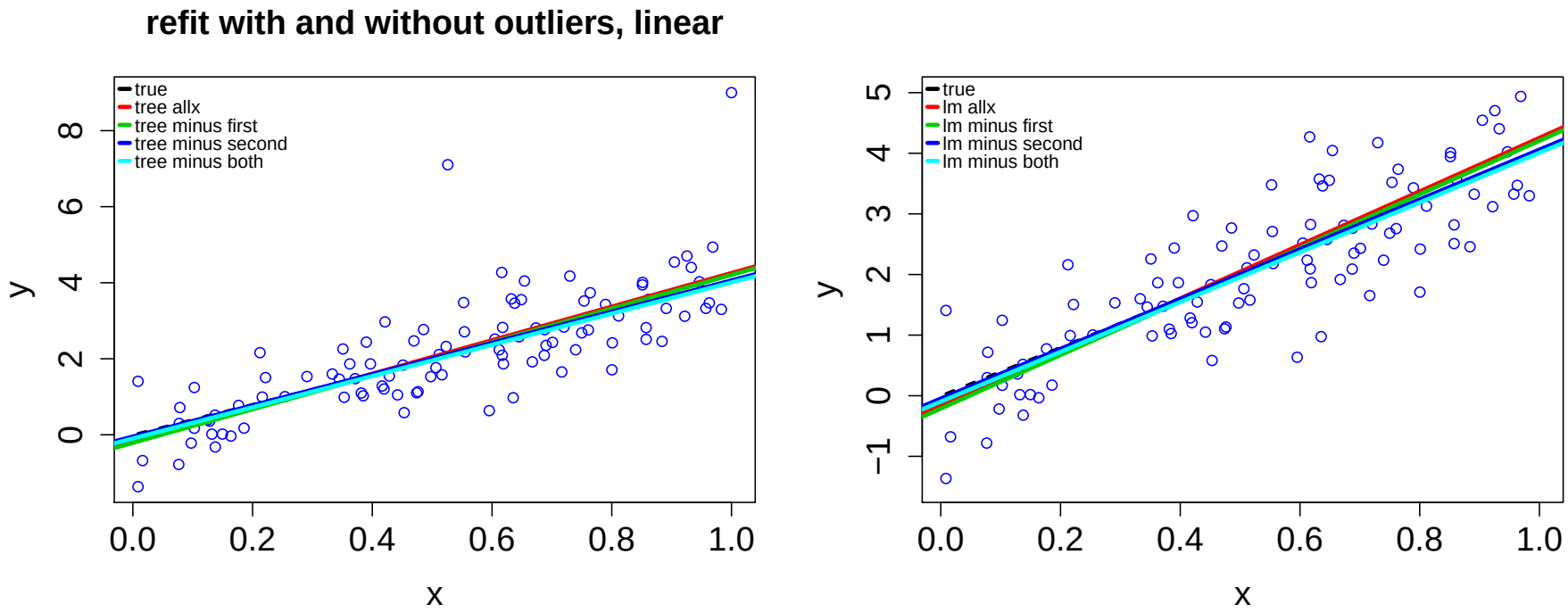
Cook's Distance (the usual linear model version).



The two outliers are identified as being potentially more influential than the other observations.
Suggests the outlier at the edge may have more influence.

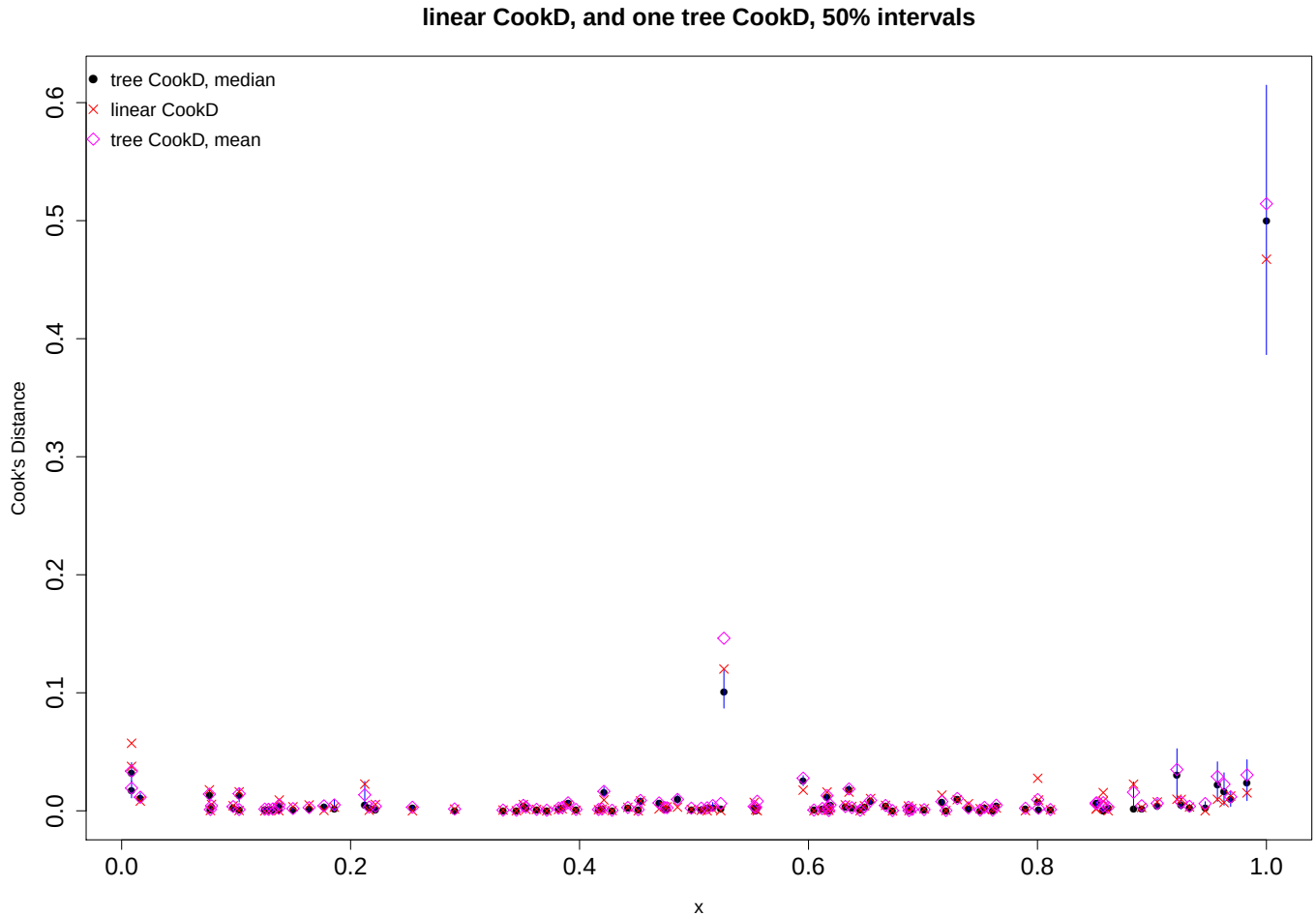
Refit the linear model without and without the outliers.

"first" outlier is the one in the middle, "second" is the one at the right edge.



The outliers are not strongly influential, but, as predicted by Cook's Distance, the one at the edge is more influential.

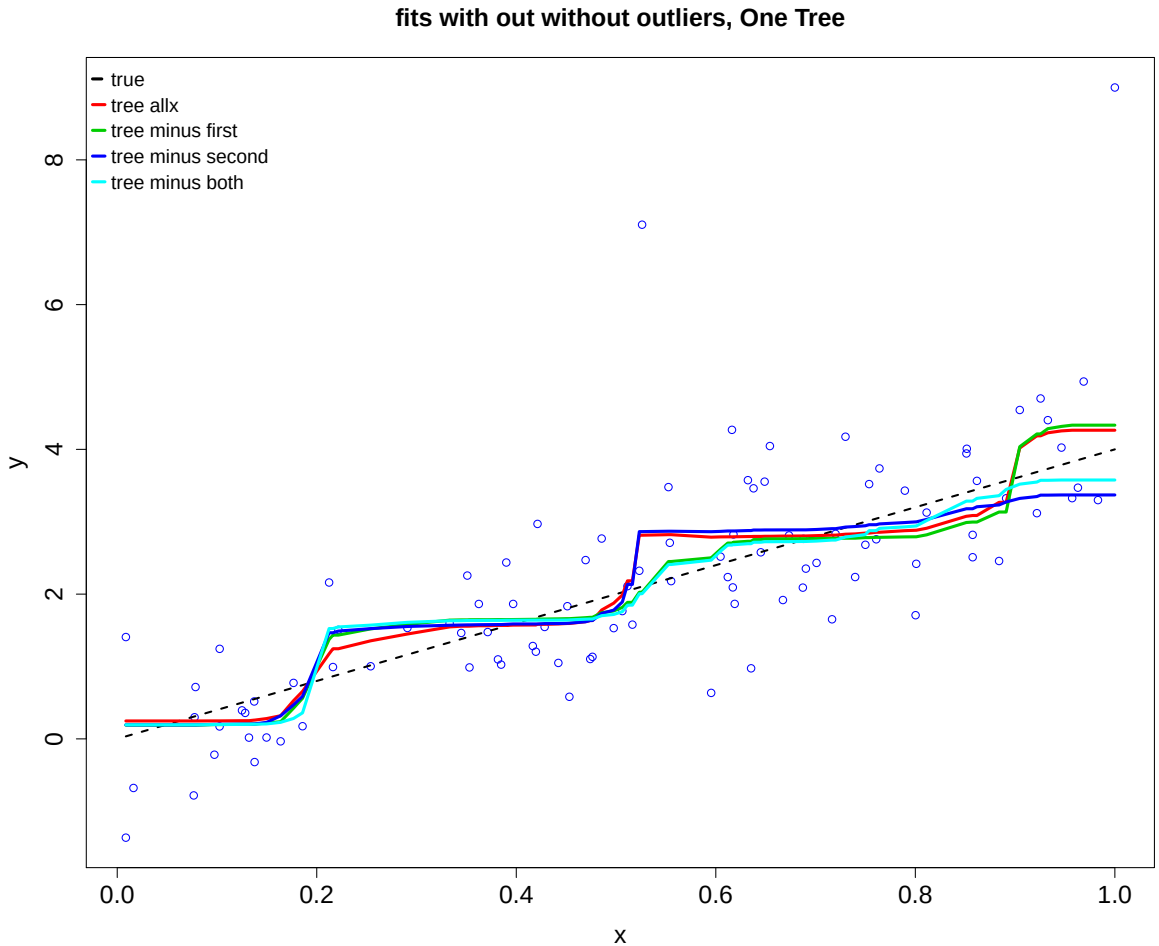
Cook's Distance, linear and the tree version.
Vertical lines indicate 50% posterior intervals for the tree Cook's distance.



Linear CookD and tree CookD very similar !!!

Refit the single tree model with and without the outliers.

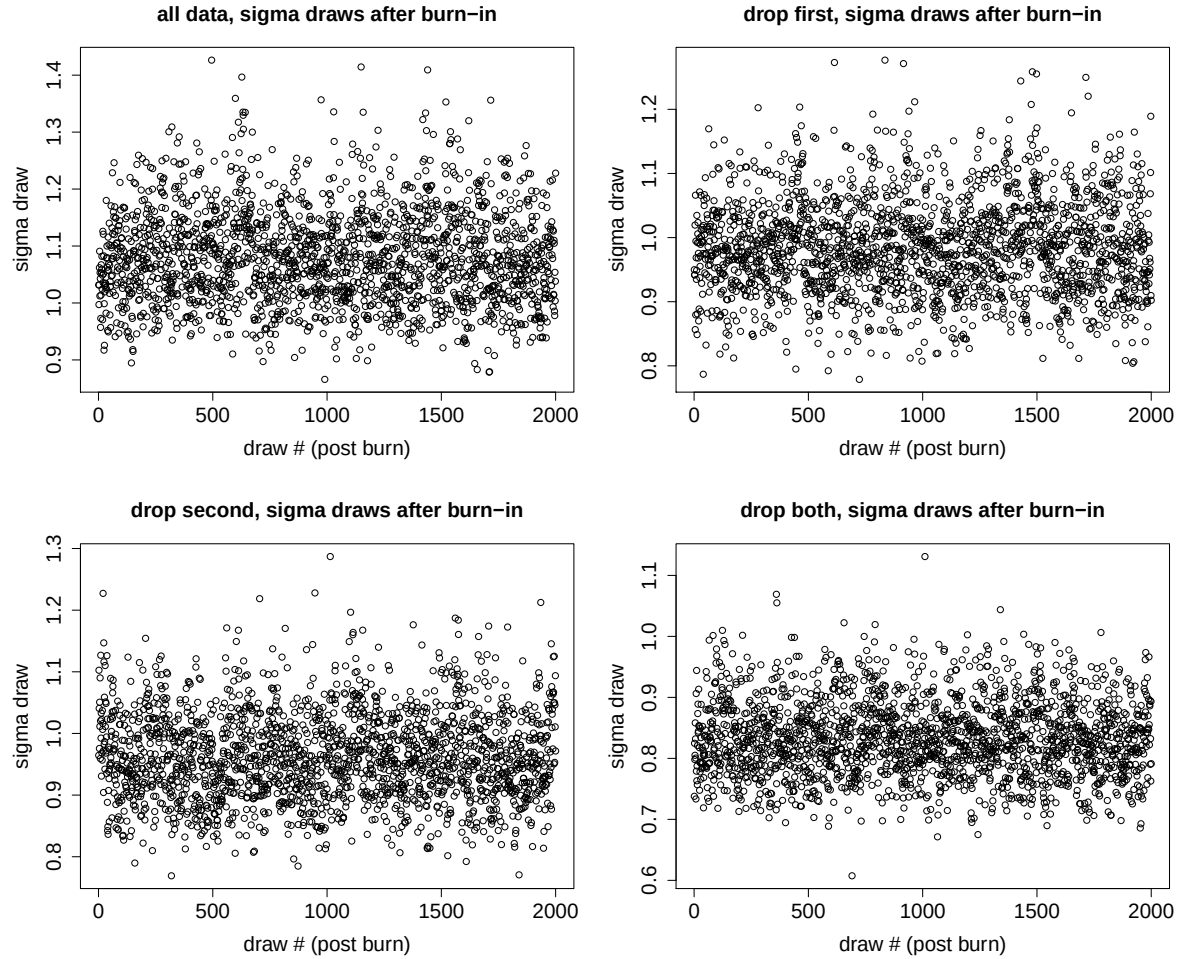
"first" outlier is the one in the middle, "second" is the one at the right edge.



Both outliers are influential!!

Does appear that the second outlier is more influential.

Check MCMC convergence!! Draws for σ (after a burn-in) from the four different data sets obtained by dropping/or not one of the two outliers.



Looks great!!

Note: this code used Pratola's enhanced single tree MCMC.

4. Cooks' Distance for A Bayesian Ensemble

In BART, "Bayesian Additive Regression Trees", (Chipman, George, and McCulloch (2010)), a Bayesian approach to ensemble tree modeling is developed.

$$y = \sum_{k=1}^{m} f_k + \epsilon$$

where each f_k is the output of a single regression tree.

Thus, the overall fit is made up from contributions of m regression trees. In application m can be hundreds or thousands.

In the MCMC we fit one tree at a time by looking and the residuals from the other trees:

$$\tilde{y}_h \equiv y - \sum_{k \neq h} f_k = f_h + \epsilon.$$

We can then write f_h as a dummied regression giving

$$\tilde{y}_h = \sum_{j=1}^{B_h} I_{jh}\, \mu_{jh} + \epsilon$$

We compute a Cook's distance from this fit exactly as we did in single tree model and then average the results from the m trees.

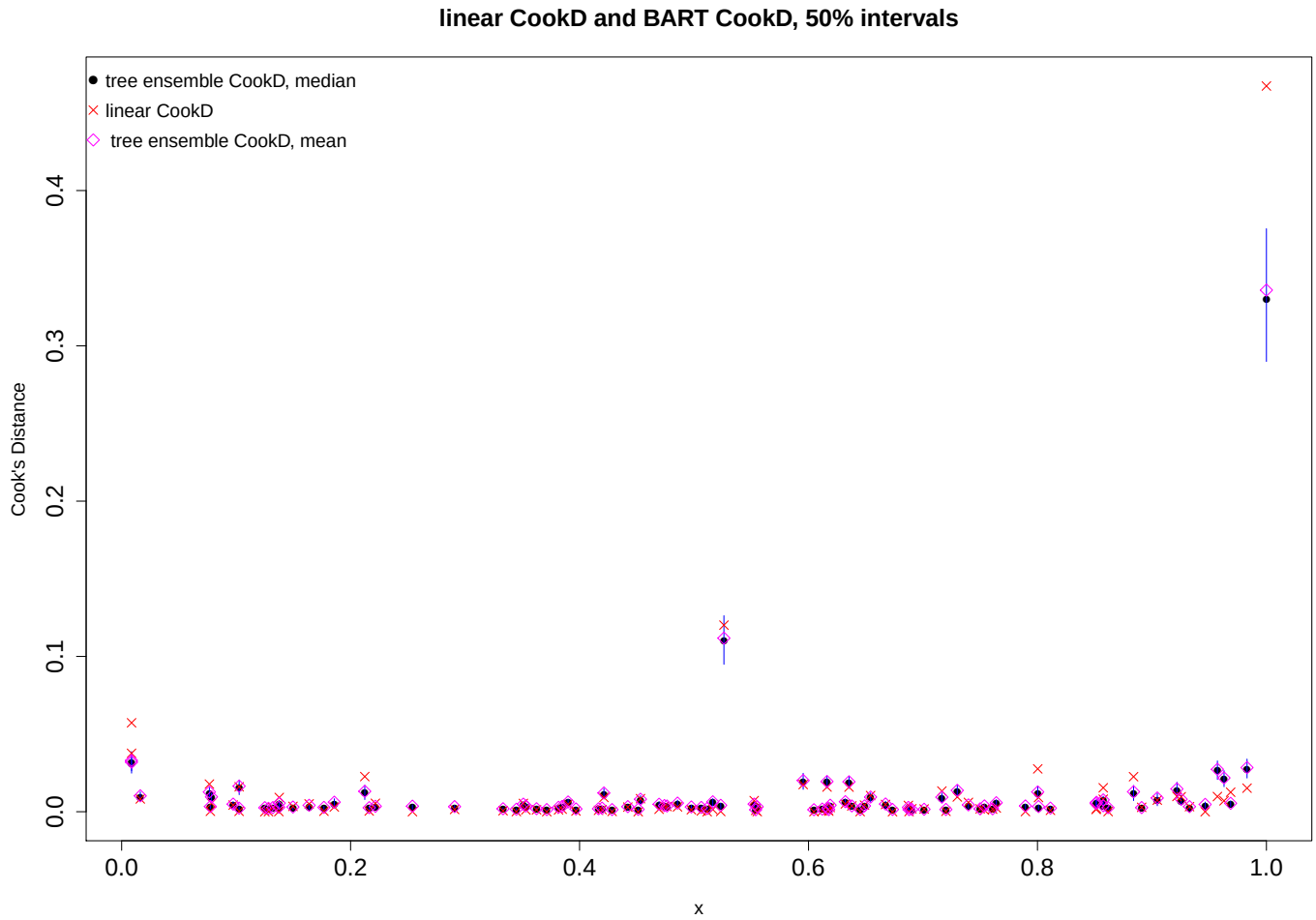
5. Ensemble of Trees: A Simple Simulated Example in 1D

We use the same simulated example we used for the single tree model.

Again, we get an average Cook's distance for every MCMC iteration.

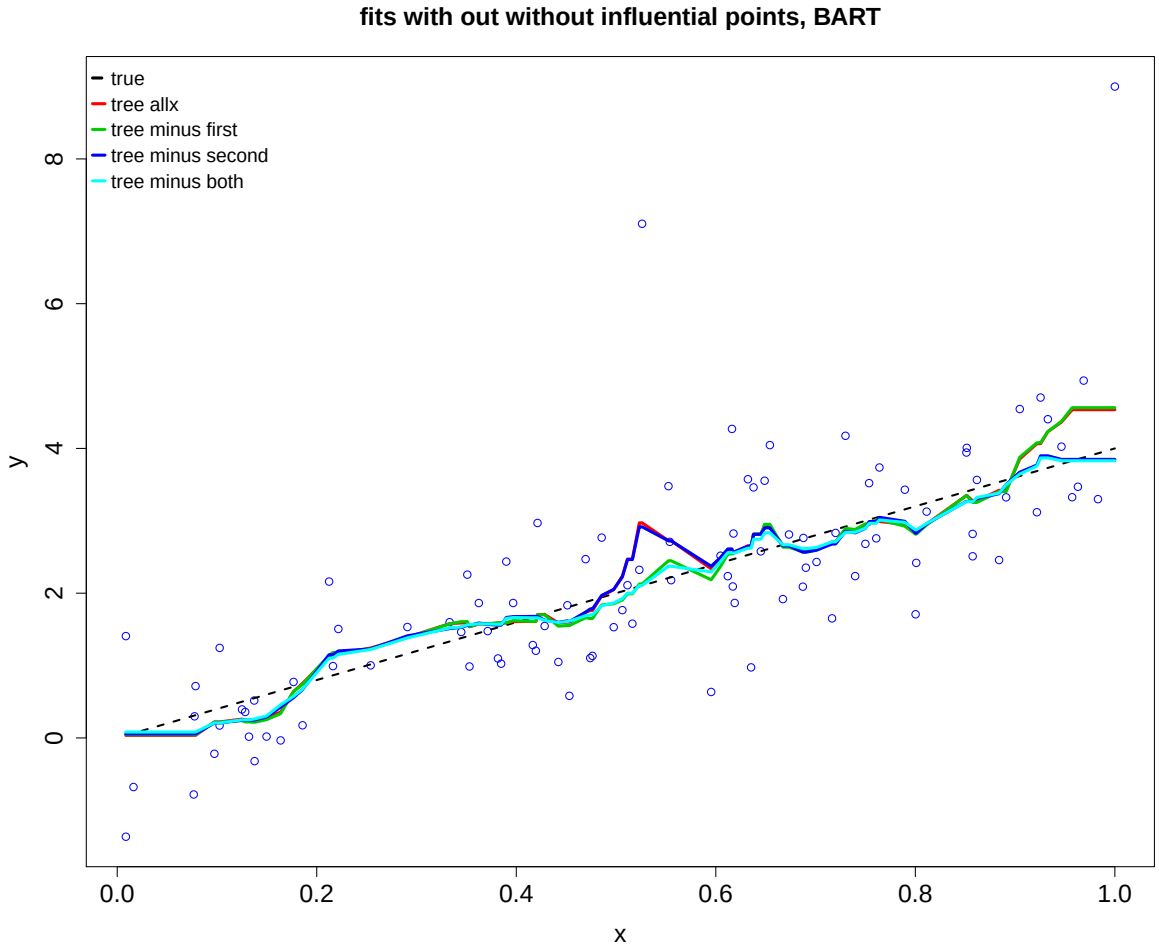
This time we use the Cook's distance averaged over the trees and refit the BART model.

Cook's Distance, linear and the BART version.



Suggestion that the outlier at the edge is not as special in BART (as compared to linear or the single tree model).

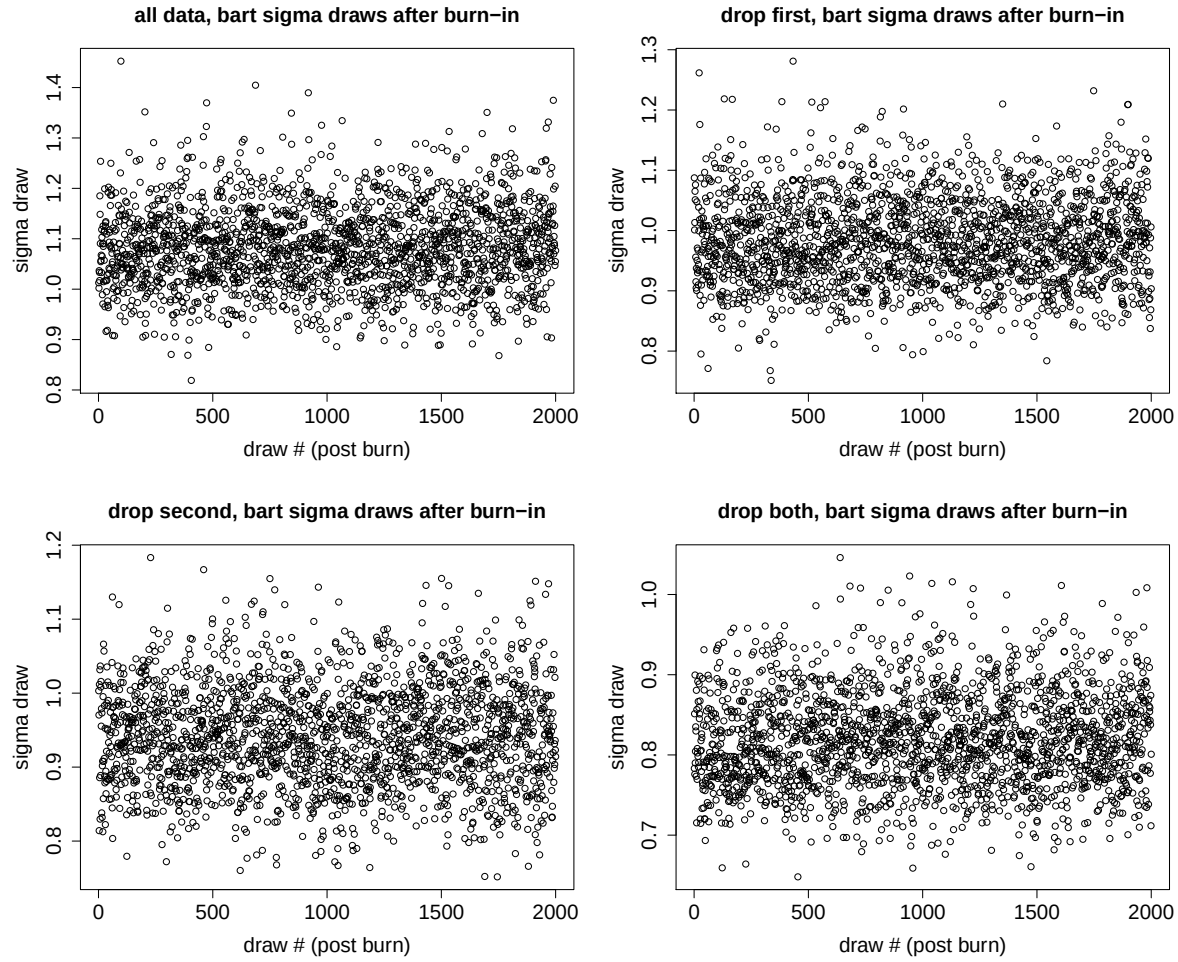
Refit the single tree model without and without the outliers.



Both outliers are influential!!

But, the influence is "local" and the outlier at the edge is not as different as it is in the linear case.

Check MCMC convergence!!



6. Cars Data: A Higher-Dimensional Real Data Example

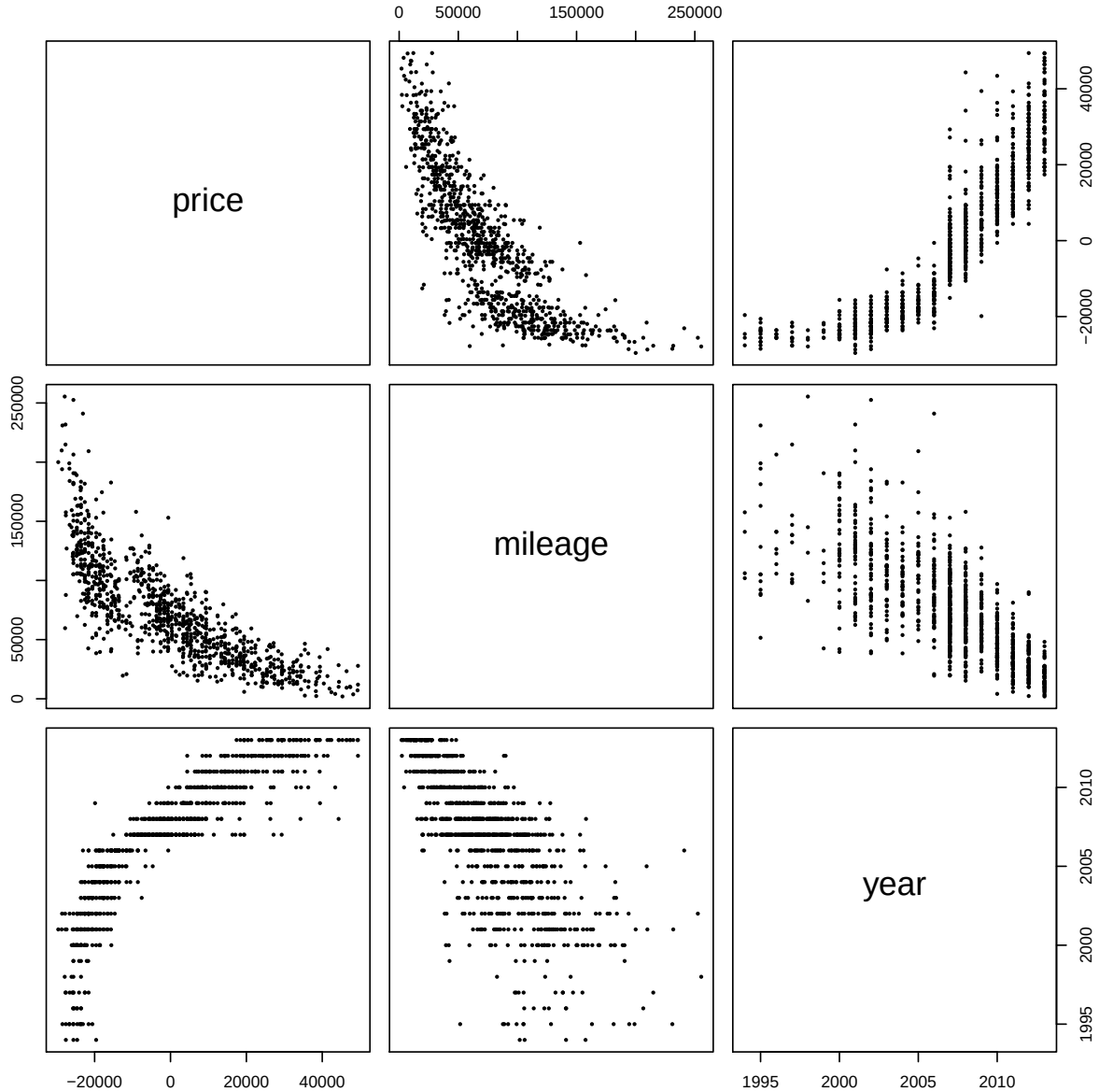
Used car sales in Texas, 15 predictors, n=1000.

Response: `price`.

Continuous predictors: `mileage`, `year`.

Categorical predictors: `trim` (430,500,550,other), `color` (black,silver,white,other), `displacement` (4.6,5.5,other), `isOneOwner` (true,false).

Spans great recession. Evidence of heteroscedasticity (Pratola et al., submitted)



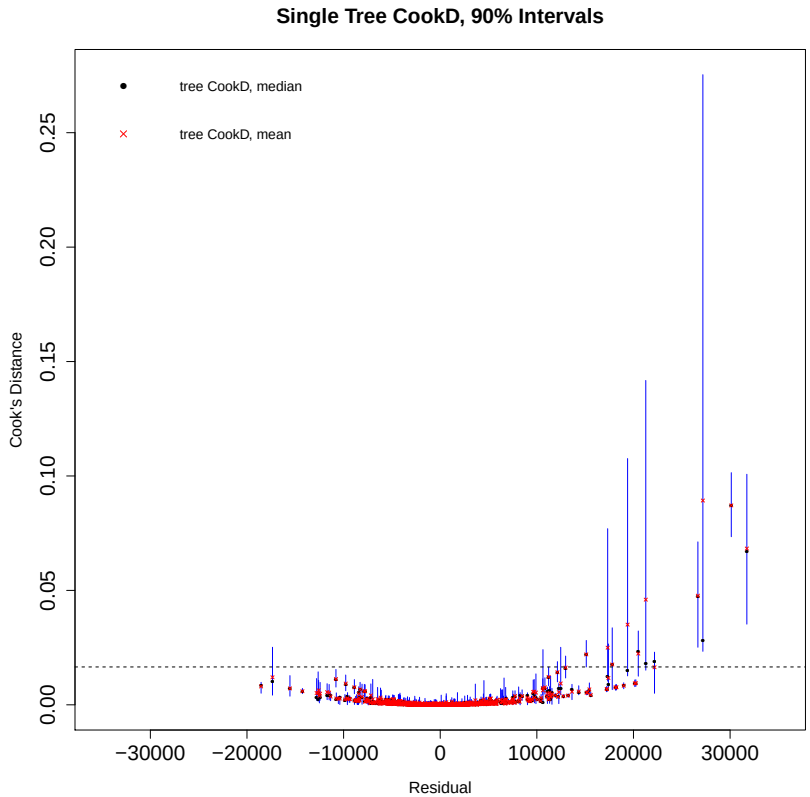
Scatterplot of continuous vars.

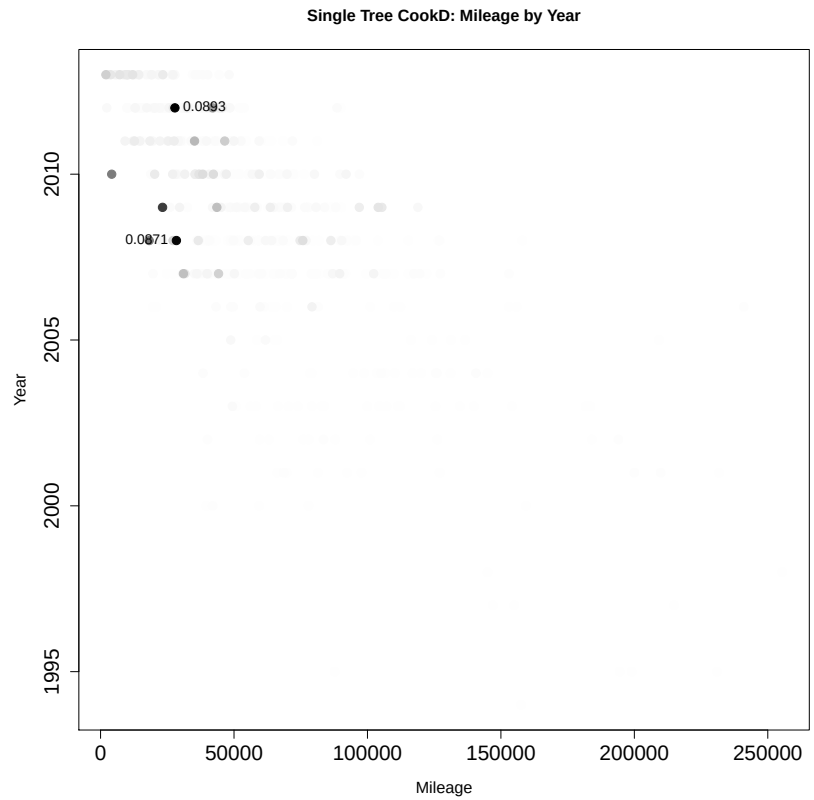
Single-tree model:

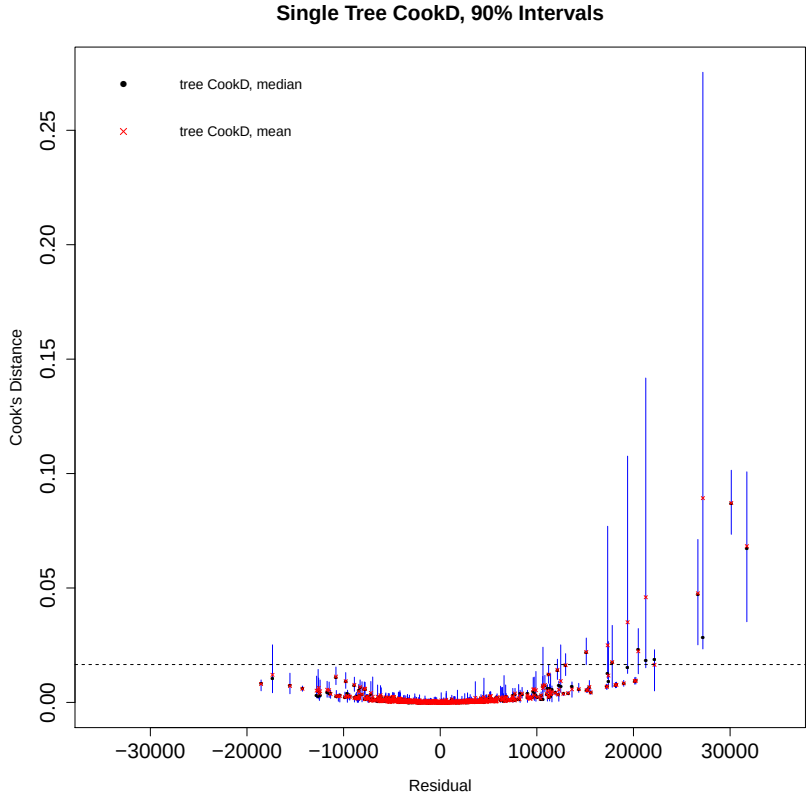
Left:
Posterior mean of residuals vs. distribution of CookD.
Dotted line represents 99th-percentile of marginal.

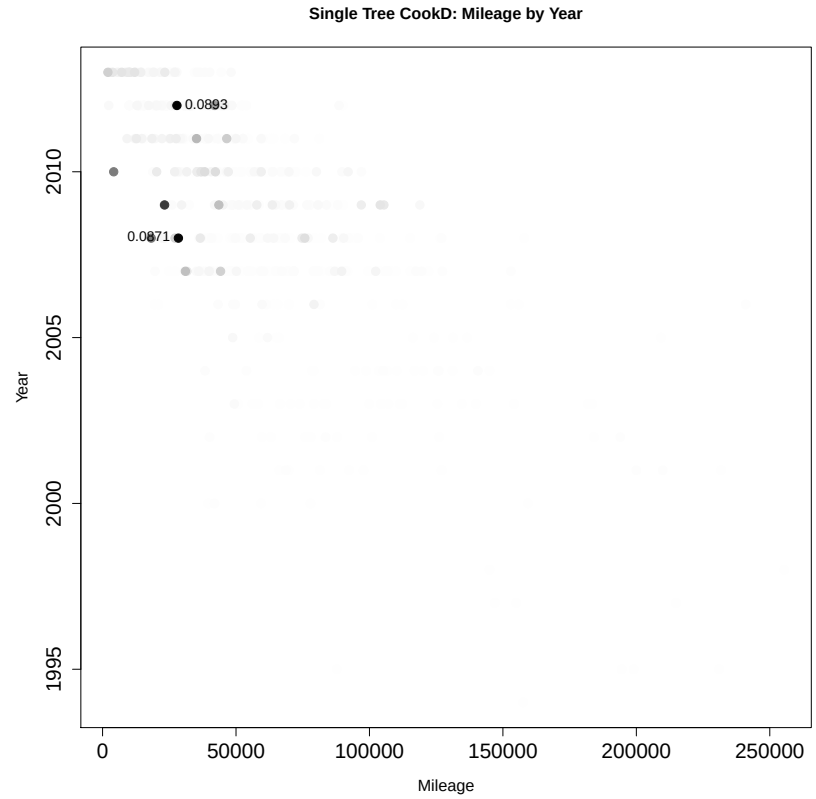
Right:

Mileage vs. Year with plotting symbol indicating posterior mean of CookD.









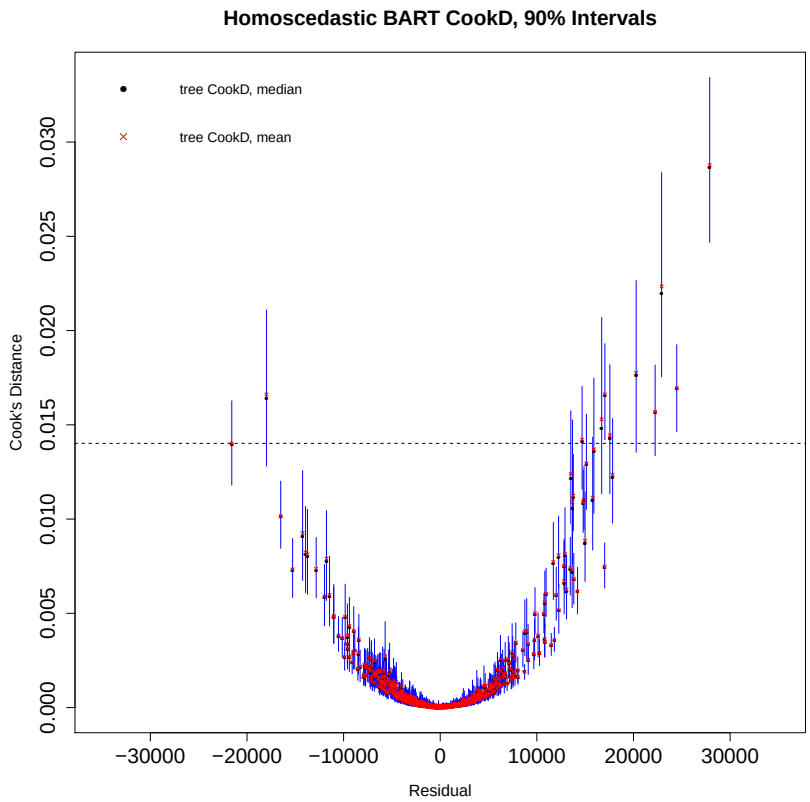
A subset of strongly influential observations!!

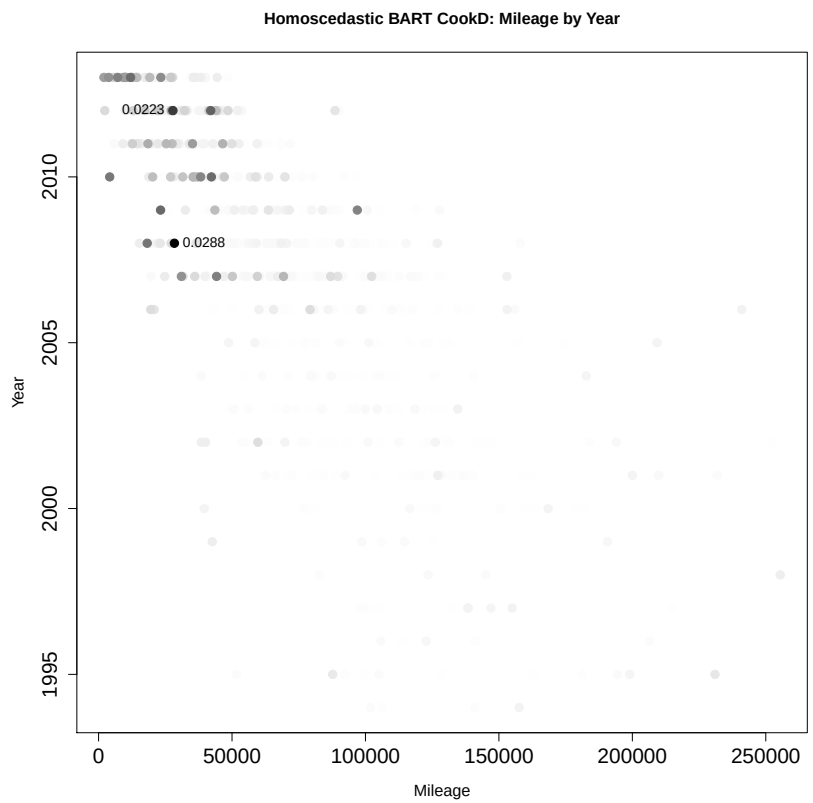
Mostly the large influence observations are the large residual ones, but there is a bit of and "edge effect" in that observations at the edges of the (Mileage, Year) plot are influential.

Influentials are low `mileage`, recent `year`.

Also `isOneOwner`, `trim` 550 or other, `color` black or silver and `displacement` 5.5 or other.

BART model:





Improved variability explained.

Still some influentials!! but less pronounced edge effect.

Again low `mileage`, recent `year`. Still `isOneOwner`, `trim` 550 or other, `color` black or silver and `displacement` 5.5 or other.

In both tree fits, most of the influentials have positive residuals (under-fitting the data)

There are some atypically expensive cars, but very few of them.

Fitting a more flexible model (BART) makes the problem more localized, *but it's still there!!*

7. Conclusion

For our Bayesian tree based models:

- Outliers can be influential. Because of the adaptiveness of the model, they may be *more* influential than in the linear case.
- Influence is more local (especially in the ensemble model) than in the linear model. At x far from that of the outlier, there may be little influence.
- Our simple use of Cook's distance gives a somewhat crude, but still useful indication of observations that could use special attention.

Cook's distance is going into the R-package and I would use it in application !!