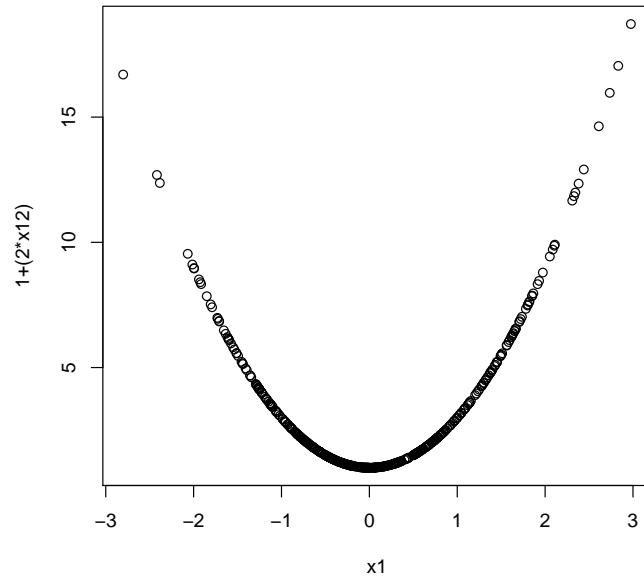


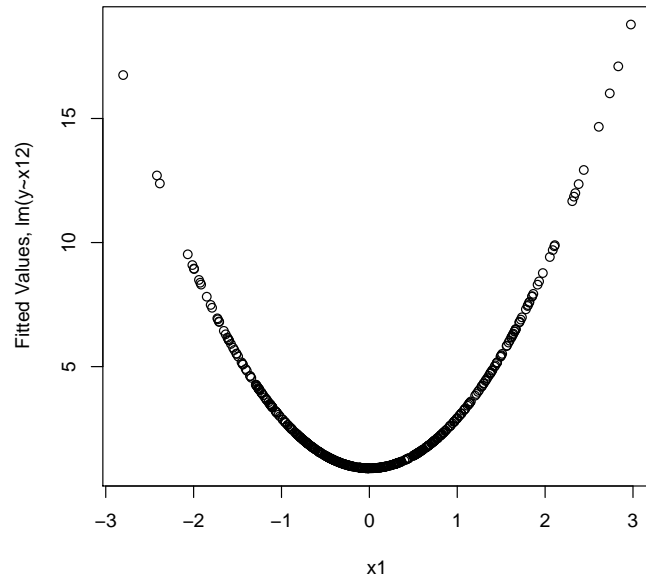
Chapter 4 Solutions

Problem Set 1

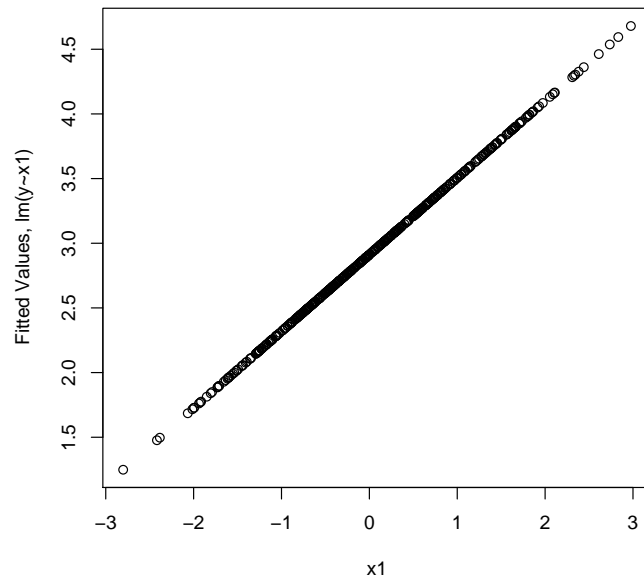
1. Below is the true $f(X)$:



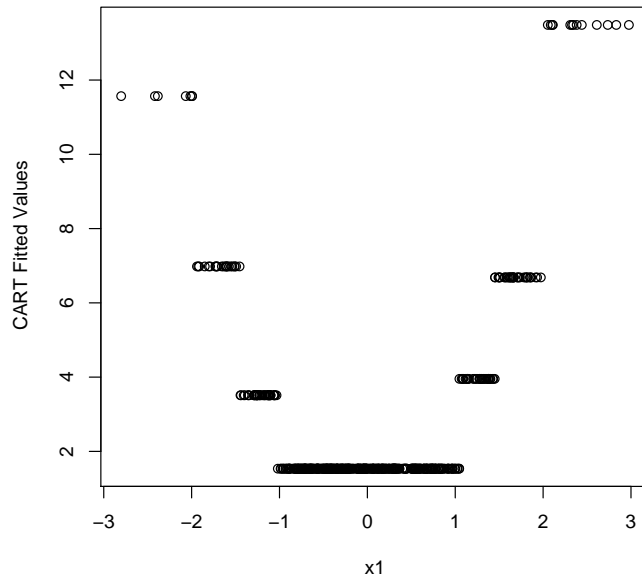
2. Below, we see that the functional form is correctly captured, as this plot and the plot from #1 look near identical.



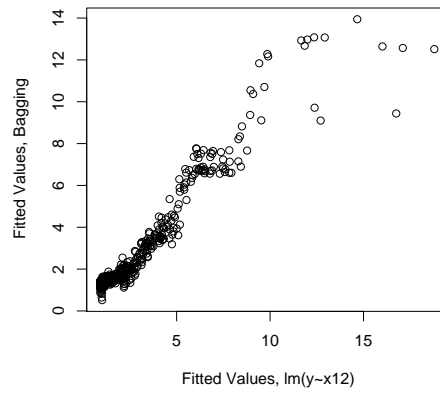
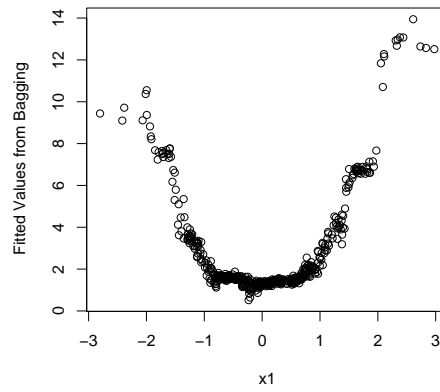
3. Given the quadratic relationship between y and x_1 , this regression model does not capture the correct $f(X)$.



4. Using CART, we no longer see a smooth relationship between the fitted values and x_1 , even though the data generation mechanism suggests the existence of a smooth relationship. Still, in comparison to the linear model in #3 above, CART does a better job of capturing the parabolic relationship between y and x_1 , even though the functional form was not specified *a priori*. Here we see that the CART model includes seven terminal nodes, each exhibiting their own predicted value.



5. The bagged fitted values are comparable to those from the linear model constructed in #2. There is a relatively smooth, parabolic relationship between the fitted values and x_1 . A scatter plot of the bagged predictions against the fitted values in #2 shows a fairly straight smoother with a slope close to 1.0. Thus, bagging does a good job of capturing the correct $f(X)$, even though the functional form was not specified *a priori*.



6. As stated in #4, CART does not show a smooth relationship between the fitted values and x_1 , even though results from #1 and #2 reveal the true conditional relationship. On the contrary, bagging is able to capture this smooth relationship between the response and predictor. Here, the extent to which bagging can smooth the conditional relationships is underscored.

Problem Set 2

1. Using CART, the root mean squared error is 758.20. Using bagging, we get a RMSE of 862.37. These are different because CART is making just one pass through the data, whereas bagging bases its results on 25 bootstrap samples (unless specified, `bagging()` in R will grow 25 trees. This can be changed easily; see the help file for `bagging()` for more details). Because out-of-bag data are used to make predictions when employing bagging, this procedure usually provides a more “honest” estimate for the root mean squared error. The results from CART tend to be overly optimistic.
2. CART produces fitted values based on a regression tree with a total of seven terminal nodes, whereas bagging produces fitted values based on many trees with varying numbers of terminal nodes. Using 25 bootstrap samples, the standard deviation of the bagged fitted values is 625.34, whereas the CART fitted values’ standard deviation is 621.90. Incidentally, if one obtains fitted values based on 200 bootstrap samples, the standard deviation is 588.83 (answers may vary slightly due to sampling variation). As more bootstrap samples are drawn, the distribution of bagged fitted values is smoothed.

Problem Set 3

1. Below are confusion tables from CART and bagging, respectively.

	Absent (Pred.)	Present (Pred.)	Model Error
Absent (Obs.)	114	19	.14
Present (Obs.)	16	63	.20
Use Error	.12	.23	Overall Error = .17

	Absent (Pred.)	Present (Pred.)	Model Error
Absent (Obs.)	106	27	.20
Present (Obs.)	27	52	.34
Use Error	.20	.34	Overall Error = .25

CART appears to outperform bagging with respect to the model, use, and overall errors. But the reality is that CART can capitalize on idiosyncratic features of the data, whereas bagging seeks to mitigate the impact of such characteristics and can ultimately yield more accurate forecasts. CART is very vulnerable to overfitting, which the averaging inherent in bagging helps to overcome. As noted earlier, however, the CART table is based on “resubstituted” data and the bagging table is based on out-of-bag data. So a CART table often gives an inappropriately optimistic representation of fit compared to bagging

2. One can see that the numbers in the cells are very similar to the numbers in the cells from the previous exercise. What we now learn is that both procedures are classifying observations in very similar ways. They do about as well with respect to each other as each alone with the data.

The numbers of observations on the off-diagonal cells are 18 and 21, respectively. Among the locations in which bagging predicted that frogs were absent, CART predicted frog presence 18 times. Among the locations in which bagging predicted frog presence, CART predicted frog absence 21 times. This means that both kinds of disagreements between the two procedures are about equally common (CART says present and bagging says absent, or the reverse).

	Absent (Bagging)	Present (Bagging)
Absent (CART)	112	18
Present (CART)	21	61