**Question 10.1**

*Using the same crime data set uscrime.txt as in Questions 8.2 and 9.1, find the best model you can using*

*(a) a regression tree model, and*

*(b) a random forest model.*

*In R, you can use the tree package or the rpart package, and the randomForest package. For each model, describe one or two qualitative takeaways you get from analyzing the results (i.e., don’t just stop when you have a good model, but interpret it too).*

1. *Regression Tree*

Here’s one possible solution. Please note that a good solution doesn’t have to try all of the possibilities in the code; they’re shown to help you learn, but they’re not necessary.

The file solution 10.1-a.R shows how to build the regression tree. It shows two different approaches: one using part of the data for training and part for testing, and one uses all of the data for training (because we have only 47 data points). A visualization of one of the trees is below.

The tree in the figure (sometimes called a “dendro gram”) shows that four factors are used in branching: Po1, Pop, NW, and LF. Notice that Pop is used in two places in the tree. Also notice that following the rightmost branches down the tree, Po1 is used twice in the branching: once at the top, and then again lower down.



It turns out that the model is overfit (see the R file). The R file shows a lot more possible models, including pruning to smaller trees, using regressions for each leaf instead of averages, etc.

The models show that Po1 is the primary branching factor, and when Po1 < 7.65, we can build a regression model with PCA that can account for about 30% of the variability. But for Po1 > 7.65, we don’t have a good linear regression model; none of the factors are significant. This shows that we would need to either try other types of models, or find new explanatory factors.

1. *Random Forest*

# The file solution 10.1-b.R shows how to run a random forest model for this same problem. In the lessons, we saw that the random forest process avoids some of the potential for overfitting. And (as the R code shows) , cross-validation shows that it works better than the previous models we’ve found for this data set.

# The R code also shows the same sort of qualitative behavior as we saw from the regression tree model. The random forest model too thinks that Po1 is the most important predictive factor for Crime. And, as we saw from the regression tree, the random forest model gives better predictive quality for data points where Po1 < 7.65 than it does for Po1 > 7.65. But unlike the regression tree, the random forest does have predictive value even for Po1 > 7.65.