Visualization and Learning Structure: Problem Session 7/8/16

Let's expand our Astrostatistics problem. Don't forget to look at the class code.

1. Download the original Hipparcos data from our website and read into R.

   `data<-read.table("HIP_star.dat",header=T)`

   - Remove the stars with missing observations (`data<-na.omit(data)`).
     You should now have 2678 stars with nine columns/variables.
     Use `attach(data)` to create a variable for each column.

   - Now calculate log luminosity (relative to the sun) as
     `logL<-(15-Vmag-5*log(Plx,base=10))/2.5`.
   - Now plot the color (B.V) against the log luminosity to double check that your picture matches the one we have been using in class: `plot(B.V,logL,pch=16)`

   Descriptions of the variables can be found on the website (Hipparcos.pdf).

2. Build a multivariate regression model predicting log luminosity from `RA, DE, pmRA, pmDE, e_Plx`, and `B.V`.
   - Check your diagnostics; do we still have the same problems?
     (When running the BoxCox, you'll need to shift `logL`: `logL.shift<-logL+2`)
   - If necessary, transform log luminosity and re-run your model.
     Re-check the diagnostics.
   - Look at a summary of your final line (`summary()`).
     Which variables have a significant relationship with log luminosity?
     Interpret the estimated coefficients ($\hat{\beta}_j$).

3. Now use a regression tree to partition the stars in similar subgroups using the same set of predictors. Use the default tree parameters/stopping criterion.
   - Which variables were chosen by the tree to be important in separating groups of stars with similar log luminosities?
   - Describe the group of stars with the lowest average log luminosity and the group with the highest log luminosities.
   - We couldn't quite isolate our gas giants today just using B.V. (color).
     Given that the gas giants exhibit the following ranges of values, was the tree able to isolate one or more groups of gas giants in its final set of leaves?

     $2.35 \leq RA \leq 358.77 \quad -78.90 \leq DE \leq 70.27 \quad -397.51 \leq pmRA \leq 760.35$
     $-473.14 \leq pmDE \leq 284.44 \quad 0.45 \leq e\_Plx \leq 1.25 \quad 0.76 \leq B.V \leq 1.23$
     $0.22 \leq logL \leq 2.11$