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

Returning to our 2010 World Cup data:

Our response variable is *PercShotsOnTarget*: the overall percentage of shots taken on target. Our four predictor variables are *ShotsExclBlocked*, *GoalsToShotsRatio*, *AvgGoalsConceded*, and *PercTacklesWon*.

More details about these variables are in the posted World Cup Handout.

Download the *2010TeamData* data set from our website and read into R (`read.table()`).

1. For each of our four predictor variables, build a regression model predicting *PercShotsOnTarget*. (e.g., `line1<-lm(ShotsExclBlocked, PercShotsOnTarget)`.

   For each model,

   - Create a residual diagnostic plotting the predicted values against the residuals (e.g., `plot(line1`fit`,`$line$`1res,pch=16)`).

     Describe whether the normal error assumptions (linear relationship, expectation zero, constant variance, normally distributed errors) look like they're met or violated.

     If you type `par(mfrow=c(2,2))` before plotting the diagnostics, all four graphs will be on one page

   - Look at the summary information and decide whether or not there is a significant linear relationship between the two variables.

2. For each of our predictor variables, experiment with fitting a LOWESS smoother by trying different window widths ($f = ...$). `help(lowess())`.

   Which f value do you think is the most appropriate and why?

3. For each of the predictor variables, experiment with fitting a cubic smoothing spline by trying different `df` values and different numbers of knots (`nknots`). `help(smooth.spline())`

   What kind of changes do you see as you experiment with knots?

4. Generally compare/contrast your locally weighted polynomial smoothers (LOWESS) to your piecewise cubic polynomial splines (smooth.spline).

   What do you think? Which do you prefer?