Visualization and Learning Structure: Problem Session 7/6/16

Data was collected on the 32 countries who participated in the 2010 World Cup. Your response variable of interest is *PercShotsOnTarget*, the overall percentage of shots taken on target. Your five possible predictor variables are *ShotsExclBlocked*, *GoalsToShotsRatio*, *AvgGoalsConceded*, *PercTacklesWon*, and *Fouls*. More details about these (and other) 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 variable, find the mean, variance, and median.

   Also create a histogram (colored, labeled, titled) with appropriate numbers of bins (e.g. `breaks=10`). *(If you type* `par(mfrow=c(3,2))` *before creating the histograms, it will put all six on the same graph page.)*

   Describe the distributions. *(centrality, skew, shape, symmetry, tails, etc)*

   Identify any potential outliers.

2. Plot each predictor variable (x-axis) against the response variable (y-axis) in a labeled, titled scatterplot. Use different `pch` values.
   (five graphs total; can use `par(mfrow=c(3,2))` *again)*

   Describe each bivariate relationship. Does it exist? linear? nonlinear? positive? negative? Again identify any potentially influential points.

   For each predictor variable, describe why or why not a linear regression model using each variable to predict percentage of shots on target might be appropriate (based on our model assumptions).

   Which predictor variable do you think would be the most appropriate? Why?

3. Taking a closer look at the relationship between *GoalsToShotsRatio* and *PercShotsOnTarget*, someone proposes that the relationship follows a normal error linear regression model with $\epsilon \sim N(0, 9.5^2)$ and true regression function $E[Y] = 27 + 1.25X$.

   (a) Write down the theoretical regression model in context; include assumptions.
   (b) Interpret the proposed β_0, β_1 values in context.
   (c) Add the proposed function to a scatterplot of *GoalsToShotsRatio* vs. *PercShotsOnTarget* (`abline(27,1.25,lwd = 2,col=3)`.
   (d) Do you agree with the normal error regression model assumption? Why (not)?

4. In class, we saw two sets of error assumptions for the simple linear regression model differing only in the assumption of a normal distribution for the stricter model. In (3), a linear regression model was proposed for the relationship between *GoalsToShotsRatio* and *PercShotsOnTarget*. We'll now use the same underlying regression function to generate three sets of *PercShotsOnTarget* response values using different error distributions with expectation zero, constant variance.

   Using R, create the following three graphs.

   Graph 1:

   - Use the *GoalsToShotsRatio* variable as our independent predictor variable X.
   - Using our X-values, generate the corresponding 32 Y-values from the regression model: $Y_i = 27 + 1.25X_i + \epsilon_i$ where $\epsilon_i \sim Unif[-9.5, 9.5]$ (see `help(runif())`)
   - Plot the (X_i, Y_i) for $i = 1, 2, 3, ..., 32$ as black filled-in circles (`pch=16`)
   - Add the underlying regression function to the plot as a line (see 3c)
   - Find $E[Y]$ for $X = 5$ and $X^{'} = 15$ (plug into underlying regression line). Add these two points $(X, E[Y|X]); (X^{'}, E[Y|X^{'}])$ (`help(points)`) to the plot as red x's (`pch=4,col="red"`).

   Graph 2:

   same as Graph 1 except $\epsilon_i \sim t_2$, the t-distribution with two degrees of freedom (`help(rt())`)

   Graph 3:

   same as Graph 1 except $\epsilon_i \sim N(0, 9.5^2)$ (`help(rnorm())`)

   Compare and contrast the three graphs.
   Include explanations as to **why** they are similar or different.

   Looking back at your scatterplot from (3), do any of the error distributions seem like a good match for the true *GoalsToShotsRatio*, *PercShotsOnTarget* values? Why/why not?