Week 8 Regression Clinic

Instructions: Your instructor will supply an R dataset (with the file name london.Rdata) which you can read directly into R and which will appear in the working environment as a dataframe object entitled “london.” This data frame contains a list of n=203 countries that participated in the 2012 summer Olympics in London. In addition to the country name, each row contains the number of gold, silver, and bronze medals, as well as a measure of average personal income (calibrated in $10,000s) and a measure of overall population size (calibrated in billions of people). The research question is to ascertain whether a country’s wealth and population size can predict its Olympic medal attainment.

This clinic unfolds in three phases: 1) creating a measure of medal attainment; 2) resolving skewness of measures; and 3) creating predictive models.

**Phase 1 Instructions**: Read in the London dataset and examine it. Each row contains the number of gold, silver, and bronze medals. This phase focuses on combining those three counts into a single measure of Olympic achievement.

1. Place the london.Rdata file in a convenient location on your computer. Read it in using the open dialog under the Environment tab in the upper right pane of R-Studio. The icon looks like a tiny yellow folder with a little green arrow. When you are successful, you will have a new data frame object, called “london,” in your working environment. Type View(london) at the command line and review the data.
2. Run hist(london$Goldmedals) and review the shape of the distribution. Also run histograms on the other two types of medals. Run summary(london$Goldmedals) and paste in the results below. Also run summaries on the other two types of medals and paste in the results below.
3. There are many different ways of combining count measurements such as the ones you have just reviewed. Here are three of them:  
     
   \* Simple sum: Add together the gold, silver, and bronze medals to create a count of the total medals. Advantages: easy to understand. Disadvantages: does not give any additional weight to gold or silver medals.  
     
   \* Borda count: 18th century mathematician Jean-Charles de Borda developed this voting system, which gives higher weight to the first ranked object, etc. In our case, every gold medal would count for 3, every silver for 2, and every bronze for 1.  
     
   \* Dowdall system: 20th century politician Desmond Dowdall created a variation on the Borda count which treats the first rank as 1 and then reduces the weight of every subsequent rank. In our case, every gold medal would count for 1, every silver for 1/2, and every bronze for 1/3.  
     
   Calculate a new variable for each of these methods, using the three columns of medal data from the london data frame. Run summary() on each of your new variables and paste them below.
4. Create a new data frame that contains the following fields:  
     
   The name of the country: london$Country  
   The simple sum of medals  
   The Borda count of medals  
   The Dowdall system measure of medals  
   The income for the country: london$Income

The population for the country: london$Popnsize  
  
Run summary() on your new data frame and past the results below.

1. Run a correlation matrix on the numeric values in your new data set. Hint: You can’t run cor() on the whole data frame because of the text data in the first column. One way to get around this is to remove the first column from the analysis with the square brackets notation. I called my new data frame newLondon, so my code would look like this:  
     
   cor(newLondon[,-1])  
     
   Paste in the correlation matrix below. Based on the strength of the relation between each of your new medal count variables and the two predictors (Income and Popnsize), which of the medal count methods do you think works best?

**Phase 2 Instructions**: Regular linear regression with ordinary least squares fitting expects the input variables to be normally distributed. Both the predictors and the outcome variables in this data set are substantially positively skewed (long right-hand tail). We should try one or more transformations that reduce the skewness of these variables.

1. First, let’s document the skewness of the variables. The e1071 package contains a skewness() procedure:  
     
   install.packages(“e1071”)  
   library(e1071)  
   skewness(london$Income)  
     
   Run the skewness() command on all of the variables of interest. Also run a histogram on each one so that you can see the shape that is associated with strong positive skewness.
2. Now pick one of the predictor variables and try testing some different transformations to see if you can reduce the skewness. Here are some possible transformations:  
     
   sqrt() – The square root of the value  
   log() – The natural log of the value  
   asin() – Arc sine, the inverse of the sine function  
   atan() – Arc tangent, the inverse of the tangent function  
   1/X – The reciprocal of the value  
     
   For each transformation, review the result with a histogram and with the skewness() function to see if there has been improvement. Use the summary() command to review the status of the transformed variable and ensure that all is well.
3. Apply the most appropriate transformation to all of the numeric variables in the new dataset. Review the dataset with View(), summary(), and cor(). Has there been any change in the correlations between the medal variables and the predictor variables?

**Phase 3 Instructions**: Complete a regression analysis of the transformed Olympic medal data. Choose the dependent variable (simple count, Borda, or Dowdall) that you believe to be optimal, based on your work on Phase 1.

1. Create an lm() model that predicts your outcome variable using Income and PopnSize. Copy and paste the output below. Write a brief statement documenting the R-squared value, whether it was significant, the B-weights, whether they were significant, and anything else of interest in the output.
2. Create an lmBF() model with at least 10,000 posterior estimates. As before, predict your outcome variable using Income and PopnSize. Copy and paste the output below. Write a brief statement that reviews the HDIs for all relevant parameters.
3. Using the posterior distribution of sig2, calculate and plot a posterior distribution of R-squared values for the lmBF() model. Write a brief statement describing what you see in the resulting histogram.
4. Create a product term by multiplying the two predictors together. Add that product term to your new/transformed data frame. Recreate your lm() model adding the product term as a third predictor. Write a brief statement that compares the R-squared value from Question 9 with the R-squared value from this analysis.
5. Integrate all of the information from Questions 8-12 to create a unified interpretation of the results. Answer the research question: how well can a country’s wealth and population size predict its Olympic medal attainment?