MONEYBALL OLS REGRESSION PROJECT (250 Points)

This data set contains approximately 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season. You are to use OLS ("Linear") Regression and the given statistics to predict the number of wins for the team. You can only use the variables given to you (or variable that you derive from the variables provided).

DELIVERABLES:

- Your write up in PDF Format. Your write up should have four sections. Each one is described below. **(160 Points)**

- A file that contains all the R (or SAS) code you used in your analysis. I should be able to run this file and get all the output that you got.

- A stand-alone R (or SAS) Data Step that will score new data as it becomes available. This should include all of the transformations and the regression equation from your analysis. **(40 Points)**

- A CSV file which has the scored records values from MONEYBALL_TEST. There will be only two columns in this file: INDEX and P_TARGET_WINS. You will be graded on how your model performs versus my model and those of other students in the class. **(50 Points)**

WRITE UP (160 POINTS):

1. DATA EXPLORATION (40 points)

Describe the size and the variables in the MONEYBALL data set so that a manager can understand it. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

a. Mean / Standard Deviation / Median
b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed "fixed"?

2. DATA PREPARATION (40 Points)

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

a. Fix missing values (maybe with a Mean or Median value or use a decision tree)
b. Create flags to suggest if a variable was missing.
c. Transform data by putting it into buckets
d. Mathematical transforms such as log or square root
e. Combine variables (such as ratios or adding or multiplying) to create new variables

3. BUILD MODELS (40 Points)

Build at least three different LINEAR REGRESSION using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Discuss the coefficients in the model, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

4. SELECT MODELS (40 Points)

Decide on the criteria for selecting the "Best Model". Will you use a metric such as Adjusted R-Square or AIC? Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model.

STAND ALONE SCORING PROGRAM (40 POINTS)

WRITE MODEL DEPLOYMENT CODE (40 Points)

Write a Stand Alone data step that will score new data and predict the number of wins. The variable with the Predicted number of Wins should be named:

P_TARGET_WINS

The data step will need to include:

a. All the variable transformations such as fixing missing values
b. The regression formula

SCORED DATA FILE (50 POINTS)

SCORE THE MONEYBALL_TEST DATA SET (50 Points)

Use the stand alone program that you wrote in the previous section. Score the data file MONEYBALL_TEST. Create a file that has only TWO variables for each record:

INDEX
P_TARGET_WINS

The first variable, INDEX, will allow me to match my grading key to your predicted value. If I cannot do this, you won't get a grade. So please include this value. The second value, P_TARGET_WINS is the number of wins you believe the team will have in season based upon the data given to you.

Your values will be compared against …

- A Perfect Model
- Instructor's Model
- Performance of Other Students
- Predict the Average value for everybody (MEAN)
- Random Model
- Worst Possible Model

If your model is not better than simply using an AVERAGE value, you will receive negative points
If your model is not better than generating a RANDOM value, you will receive a LOT of negative points
If your model is not better than the WORST model, then it will be a WHOLE LOT of negative points.

BINGO BONUS:

If you want Bingo Bonus Points, write a brief section at the top of your Write Up document and tell me exactly what you did and how many points you are attempting. If I cannot see your Bingo Bonus work, I cannot give you credit. Bingo Bonus is difficult to grade and I don't have time to go back looking for it. If you don't tell me it's there, I cannot give you points.

The policy with Bingo Bonus is: ***All Sales are Final !***

- (20 Points) Once you select a champion model in Step 4, use PROC GLM and PROC GENMOD to do the OLS Regression. Are the results the same? Are there any differences?
  {Note: PROC GLM and PROC GENMOD are SAS procedures. The equivalent (or near equivalent) in R packages are **lm** (lm for SAS's GLM) and **glm** (glm for SAS's GENMOD).
- (20 Points) Use decision tree software such as Angoss or Weka or something else for variable selection or missing value imputation (the more use you make of decision trees, the more points you will receive). Be sure to carefully present your decision tree output so that I can see what you did.
- (20 Points) Recreate as much of the program as you can in a second program (e.g., R and SAS)
- (?? Points) Roll the dice … think of something creative and run with it. I might give you points.

Note there is a 50 point maximum for bonus points (per Unit assignment).

PENALTY BOX

- (Lose 10 Points) If you don't have PDF format
- (Lose 10 Points) If you don't have a GOOD Introduction
- (Lose 10 Points) If you don't have a GOOD Conclusion
- (Lose 10 Points) If you don't put your NAME in the file names of any files you hand in
- (Lose 10 Points) If you don't put your NAME inside of the files you hand in
- (Lose ?? Points) For anything that I think might annoy your boss !