Baseball Regression Project

This data set contains approximately 2276 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).

1. Data Exploration

Describe the size and the variables in the BASEBALL 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 checklist 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”?

1. Data Preparation

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)

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

1. Model Building

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

Discuss the coefficients in the model, e.g. 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.

1. Select Models

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.