Baseball Player Salary Modeling (SASHELP.Baseball)

Use a native SAS data set on Baseball, this data was briefly explored during the module for this unit, but you will explore it some more in this assignment. This data is in the SASHELP library and can be accessed and manipulated by creating a new data set and using the SET statement to set it equal to the Baseball data set. Note: When using the SET statement make sure you indicate that the Baseball data set is named SASHELP.Baseball so that SAS properly understands the reference.

For this assignment you will be using these variables from the data set as a viable list of predictors to create the best model in each part:

nhits nruns nouts CrRuns CrHits CrAtBat YrMajor nAtBat nAssts

Your response variable will be: **salary.**

If deemed appropriate, transformation of variables is allowed but consider that you may only need a portion, all, or none of the variables transformed.

Part A) Using one of the variables listed identify the regression model that explains the largest percentage of variation amongst the data for salary. Interpret the p-value and correlation coefficient of your model in context. Does this model violate any assumptions? Does this model appear valid? Justify your response

Part B) Now selecting from all the variables listed create the best multivariable linear regression model to predict for salary. Interpret the p-value and correlation coefficients of your final model in context. Discuss your methodology for variable selection whether manual or otherwise and discuss the strength of your model. Does this model violate any assumptions? Does this model appear valid? Justify your response.