M2: Data Assignment

Jason Merten

Data Exploration/Preparation:

The data used in this analysis comes from two sets of data: a training dataset consisting of 2276 rows of data and test dataset consisting of 259 rows of data. These two datasets contain the same 17 columns of data, of which 6 columns contain missing values: **BATTING_SO**, **BASERUN_SB**, **BASERUN_CS**, **BATTING_HBP**, **PITCHING_SO**, and **FIELDING_DP**. Two of these columns (**BASERUN_CS** and **BATTING_HBP**) contain a large amount of missing data and were filled using a normal distribution of existing sample data. Two new columns were created based on a potential use while building the model: **BATTING_1B** (single base hits) and **SB_P** (stolen base %). I decided to transform both datasets at the same time to reduce confusion in the code while simultaneously verifying the data after the transforms.

Building Models:

For building models, I decided to create three separate models utilizing different techniques. The first model was created using a stepwise function to minimize the AIC value of the original linear model. This model was the easiest and quickest way to get a functioning model with a decent R^2 value by doing both forward and backward selection.

The second model was created using subset regression utilizing an exhaustive search and some standard options. After the subset search completed, I plotted the subsets with respect to the adjusted R^2 values for each variable to create the linear model.

The final model was manually created using trial and error by performing mixed and backward selection while trying to maximize the R^2 and adjusted R^2 values. Model3.1 was created by removing variables with a high p-values while trying to maintain a high R^2 value. Model3.2 was created by adding variables and interactions based on what may be useful to help the prediction.

Selecting a Model:

To select the most optimal model I used a few different methods and measures. The first thing I looked at were the residual, QQ, standardized residual, and leverage plots for each model to look for abnormalities in each model. The next thing I looked at were some standard statistic measures for model fit: R^2, adjusted R^2, sigma (Residual Squared Error), AIC, BIC, p-value, MSE (Mean Squared Error), and PER (Prediction Error Rate). I selected Model3.2 as the champion model due to its great performance in all of the statistic metrics compared to the other models. Below is a table of the results of this analysis:

Metric/Model	**Stepwise**	**Subset**	**Model3.1**	**Model3.2***
R^2	.435	.405	.434	.449
Adjusted R^2	.430	.402	.429	.445
Sigma	11.9	12.2	11.9	11.7
AIC	17752	17853	17758	17693
BIC	17878	17933	17896	17819
p-value	1.18e-261	7.53e-245	3.04e-259	3.39e-274
MSE	140.1206	147.5049	140.2574	136.5414
PER	.1472	.1508	.1473	.1453