Tyler Rose, Nicholas Blanton, Dylan Smith, Claire Gambacorta

Project A

1. Motivation: Our group is interested in looking for what variables affect a team’s win/loss ratio. Using a regression model we can find which variables have the greatest effects on the W/L ratio. The variable with the most effect on the W/L ratio would help give insight on what a team should focus on to improve win chances.
2. Data Description: The data used is baseball statistics for the 2018 and 2019 seasons. Our predictor variables are runs, batting average, ERA, fielding percentage, team name, and year. Our response variable is the Win/Loss ratio. Runs are the number of points scored by a particular team in a particular season. Batting average describes the ratio of hits over ‘at bats.’ ERA is the average number of runs earned for a particular team during a particular season. Fielding percentage is a measurement of defensive performance on the field. We have 60 observations pulled from SeanLahman.com

<http://www.seanlahman.com/baseball-archive/statistics/>

1. Data Exploration:

FILENAME CSV "/home/u42911087/Tyler Rose Work 2/Data/Teams (project A).csv" TERMSTR=LF;

/\*\* Import the CSV file. \*\*/

PROC IMPORT DATAFILE="/home/u42911087/Tyler Rose Work 2/Data/Teams (project A).csv"

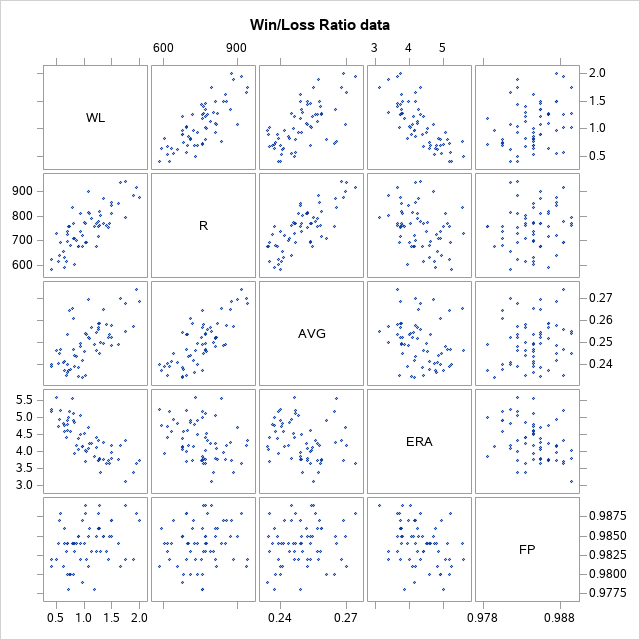
OUT=ProjectA

DBMS=CSV

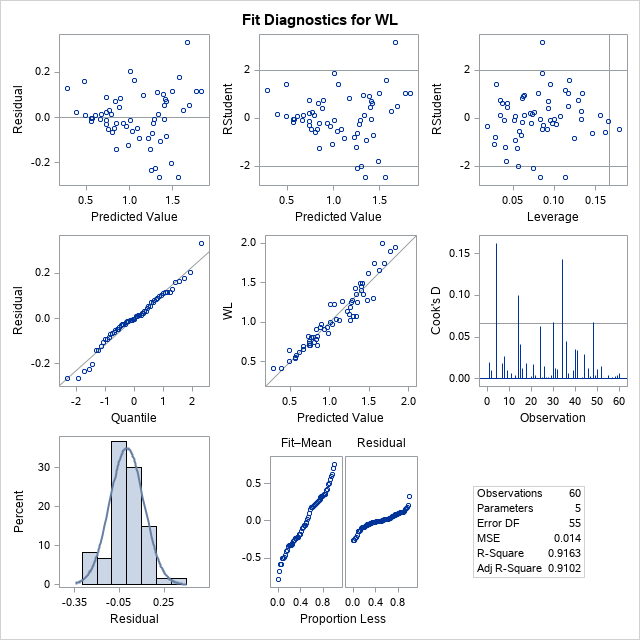
REPLACE;

RUN;

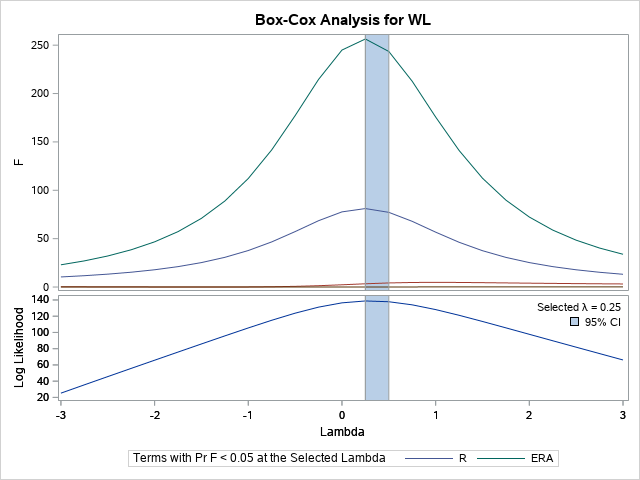
PROC SGSCATTER:



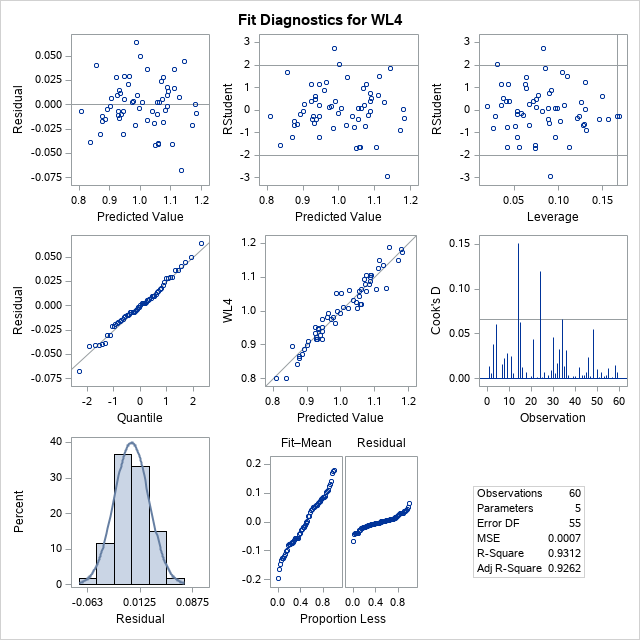
The Win/Loss ratio is positively correlated with both Runs and Batting Average, which makes sense. The ratio is inversely related to ERA which also makes sense, since ERA is a pitching statistic. Fielding percentage seems to have no relationship with any other variables, including the Win/Loss ratio.

1. Model Fitting and Analysis:
   1. Regression Modeling
      1. See SAS code
      2. R-Square 0.9163

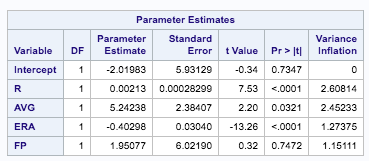
There seems to be a slight megaphone effect within the Residual vs Predicted Value plot. There are also a few outliers and one point with some concerning leverage. The QQ-plot looks pretty good though. We will perform a proc transreg to get a transformation suggestion.



* + 1. The suggested transformation is 0.25



After the transformation the megaphone effect was eliminated but the outliers and leverage concerns prevailed. Also, Batting Average went from being significant to insignificant. Because of this we think the original model is better.



* + 1. When looking into the potential interference of multicollinearity, there appears to be absolutely no signs of this being an issue. This was not expected to be an issue with every variable covering very different aspects of the game.
    2. There were a few variables that were questionable points (one with leverage and two outliers) but there was nothing that warranted removal from the model due to the points in question not having a significant effect on the data set.
    3. D

data new;

input R ERA FP WL AVG;

datalines;

700 1.5 .70 . 0.300

;

run;

data BBCombo;

set new BB3;

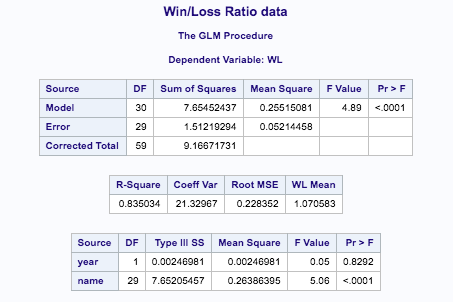
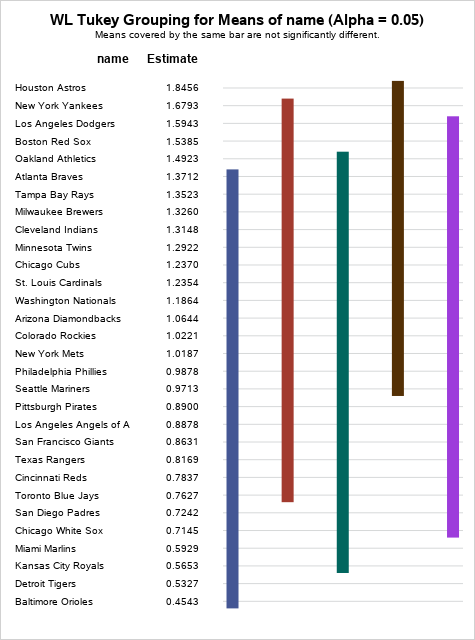
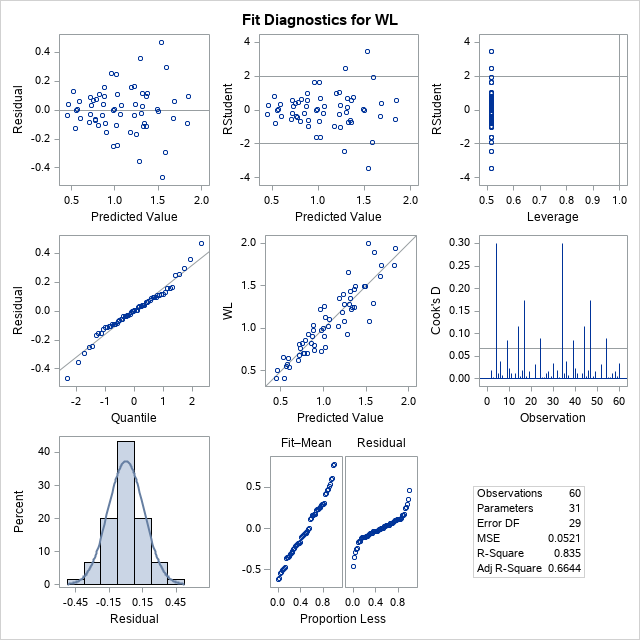
run;

proc reg data=BBCombo;

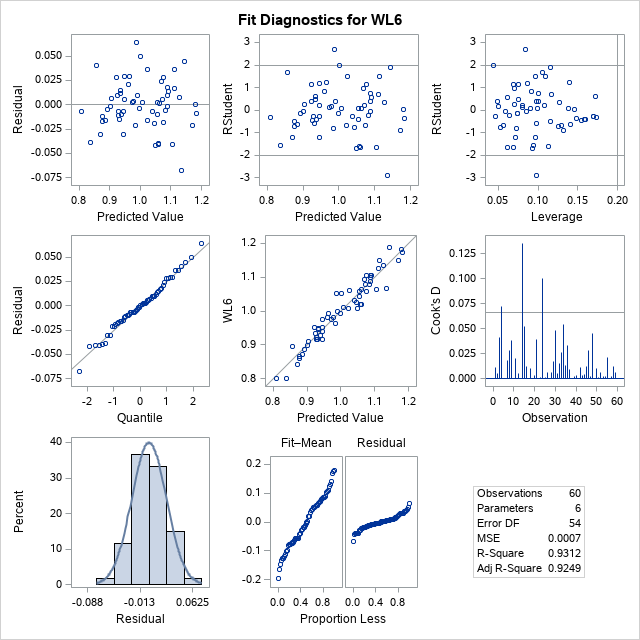
model WL = R AVG ERA FP/cli clm;

id R AVG ERA FP;

Run;

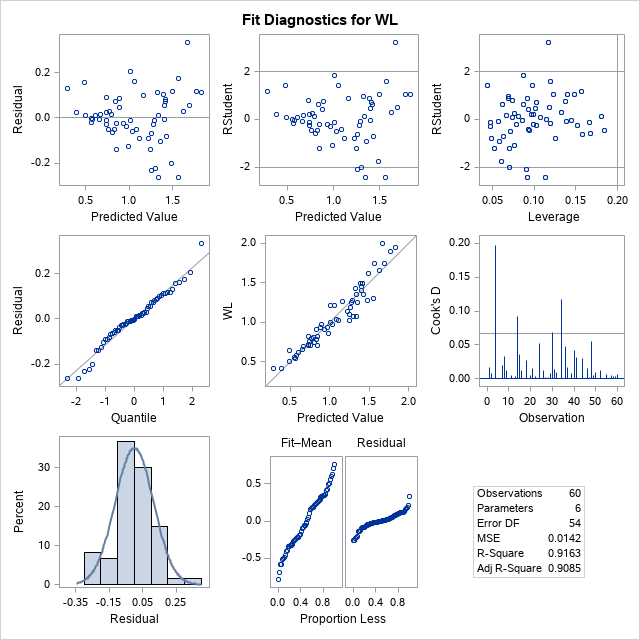
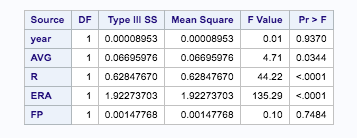
* 1. ANOVA
     1. 
     2. 
     3. 

These are the residuals for our ANOVA before any transformations. There is a serious megaphone effect in the residual vs predicted plot.



These are the residuals for our transformed ANOVA. The suggested transformation was 0.25. There was a strong megaphone effect present in the original residual vs predicted plot. This fixes that issue.

* 1. ANCOVA
     1. The only significant predictors are batting average, runs, and ERA. All three predictors have a p-value of less than 0.05.



* + 1. The residual vs predicted plot shows a possible megaphone effect, however it looks pretty good otherwise. There are a few outliers but no leverage and the QQ plot is a relatively good fit.
    2. The regression model is arguably the best fit for this data. It had the most significant variables and the highest R-Squared. Including the Quantitative and Categorical variables with the ANCOVA was an improvement over the ANOVA, the regression model still came out on top.

Conclusion: The regression model shows that runs, the earned runs average, and the batting average are significant. This suggests that these are the most important aspects for a team to focus on to have the highest chance of winning. Also, 91.6% of the error was explained by the model meaning the findings were very accurate.

We, the project team members, certify that the percentage of the effort listed by each of our names below is an accurate account of the original effort contributed by each team member in the producing of this project and report.

Nicolas Blanton 25%

Claire Gambacorta 25%

Tyler Rose 25%

Dylan Smith 25%