Stat Project

Connor Jenna Scherasade

Our Variables

WLR (y) OBP SOdivIP

RunDiff SLG HdivIF

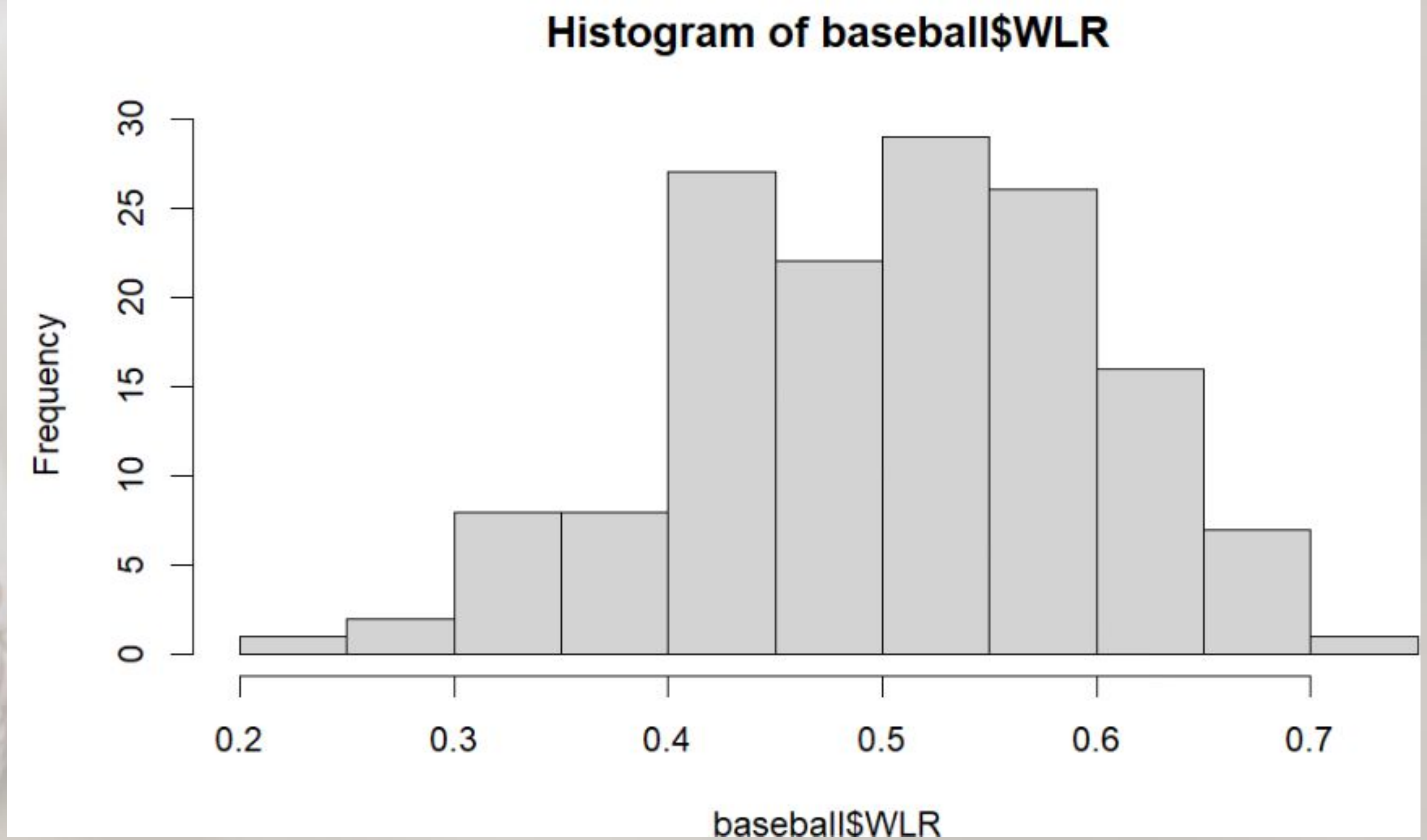
BatAge SOdivAB

PAge RAdpG

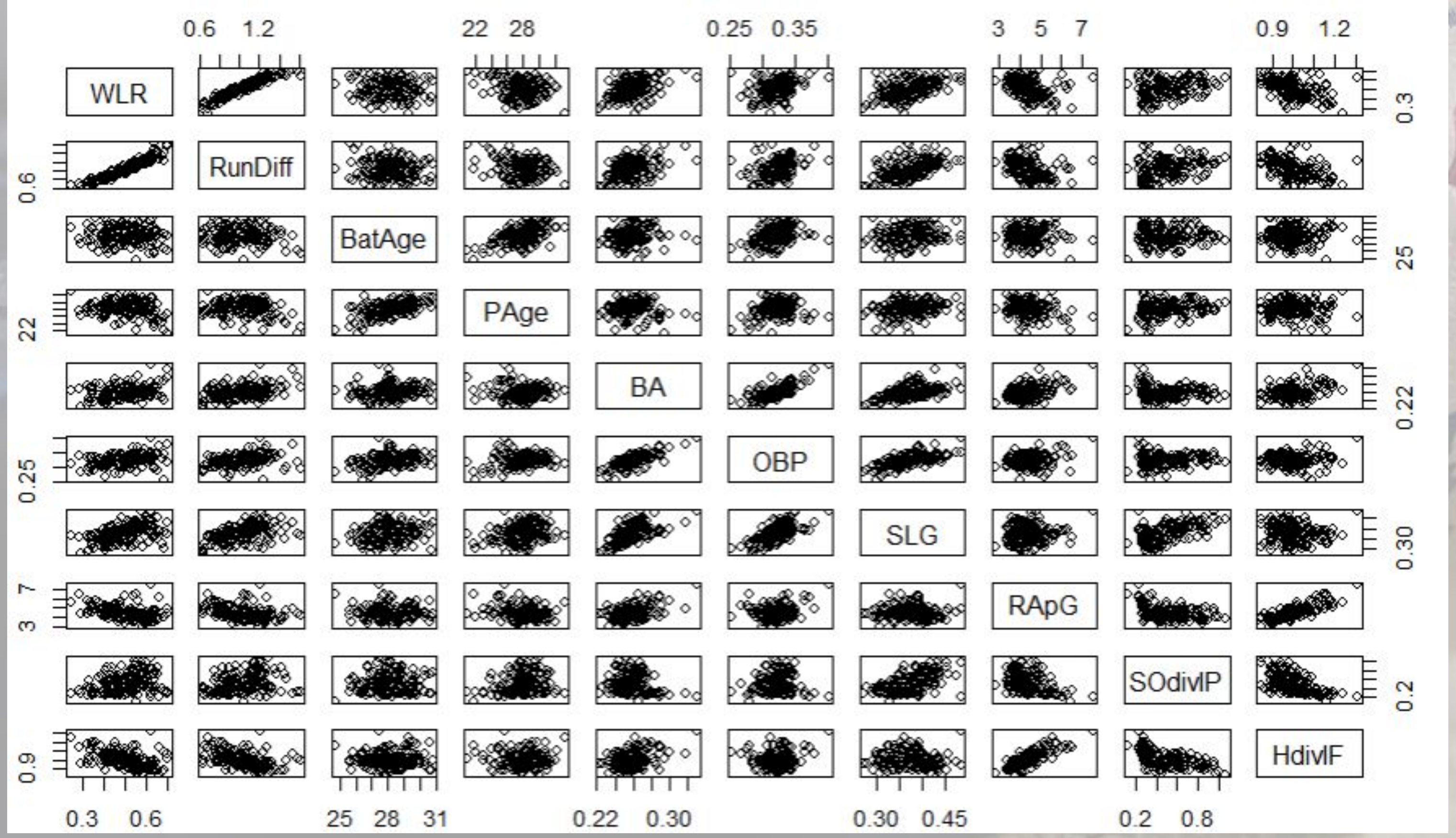
BA ERA

````
# Variable Names
```{r}
#Year = Year
#WLR = Win Loss Rate (FOR THE BRAVES!!!)
#RunDiff = Runs For / Runs Against (points scored vs points against)
#BatAge = Batter Age (avg)
#PAge = Pitcher Age (avg)
#BA = Batting Avg (times hit ball and gotten on a plate)
#OBP = On Base % (BA + walks)
#SLG = Slugging Percentage (Weight of points)
#SOdivAB (Strike Out/At Bats)
#RAdpG (Runs Allowed per Game (accounts for error in field))
#ERA (Runs Allowed per Game (by pitcher only))
#SOdivIP (Strike Outs / Innings Pitched)
#HdivIF (Hits / Innings Pitched)
```
````

The histogram distribution of WLR (Win Loss Rate), our response variable, is roughly normally distributed with a mean WLR of .51, standard deviation of .097, a minimum value of .248 and a maximum value of .705.



The scatterplot matrix below shows the correlation between the variables. Using the VIF test we eliminated multicollinear variables.



Data Exploration

Stepwise

From this point forward, we are making changes to 75% of our data (Training Data)

We used the stepwise procedure of variable selection to create our first model (baseball_1) that had only one variable, RunDiff.

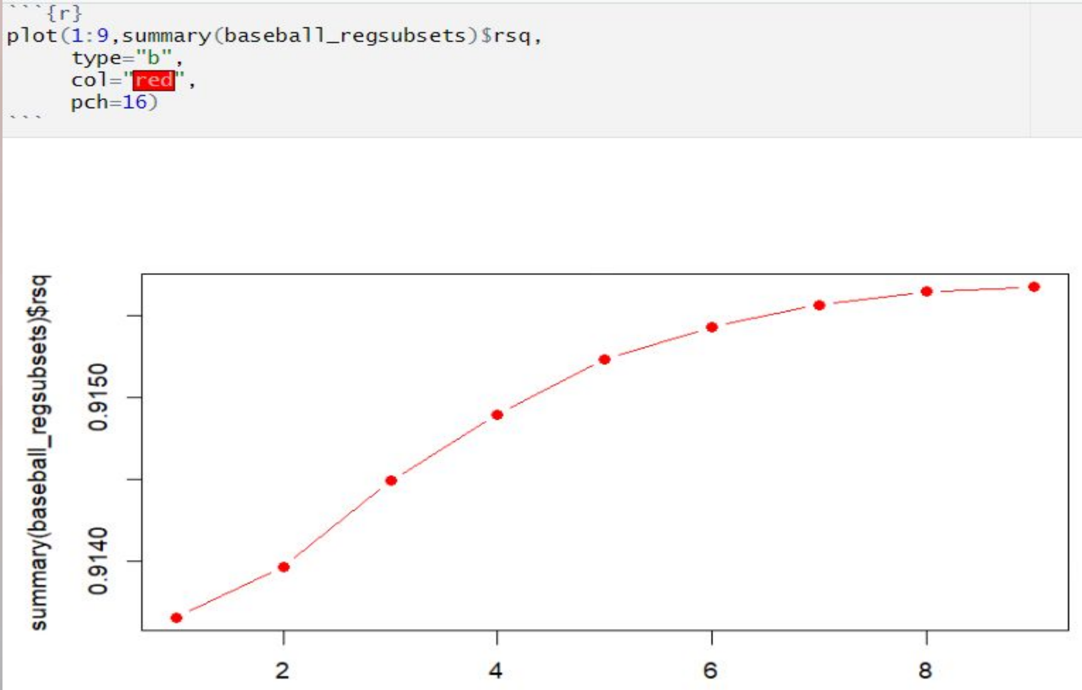
To counter any possible nonlinear-pattern, we made a second model (baseball_2) that was identical to baseball_1, but with the inclusion of another predictor variable, RunDiff^2.

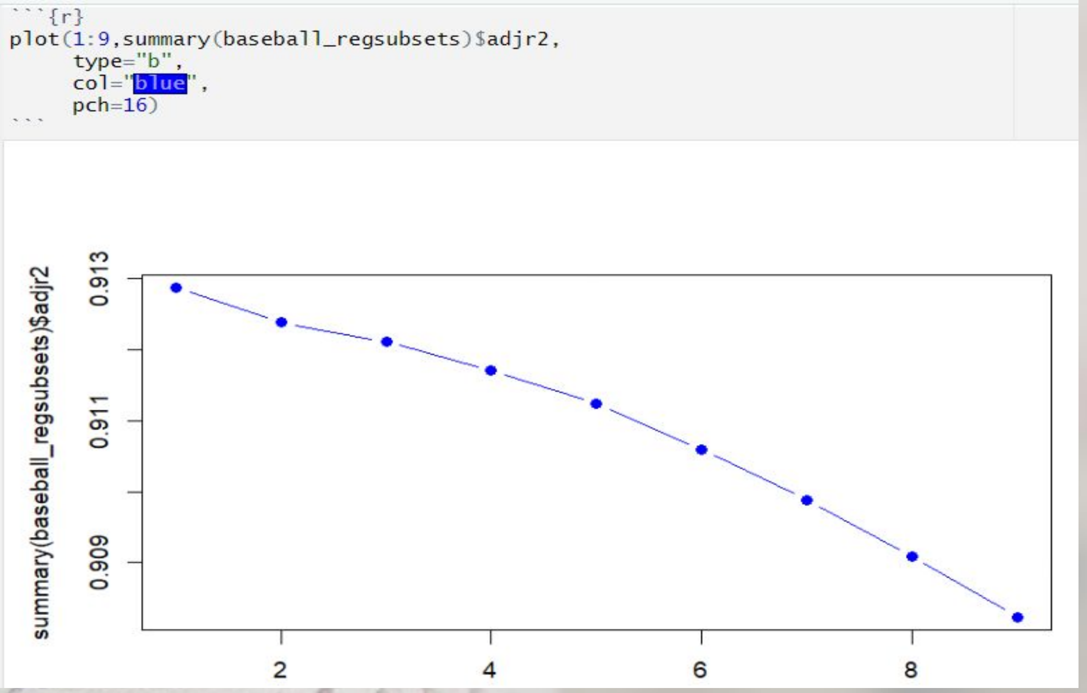
baseball_1:
Predicted WLR = .06181 + .4317(RunDiff)

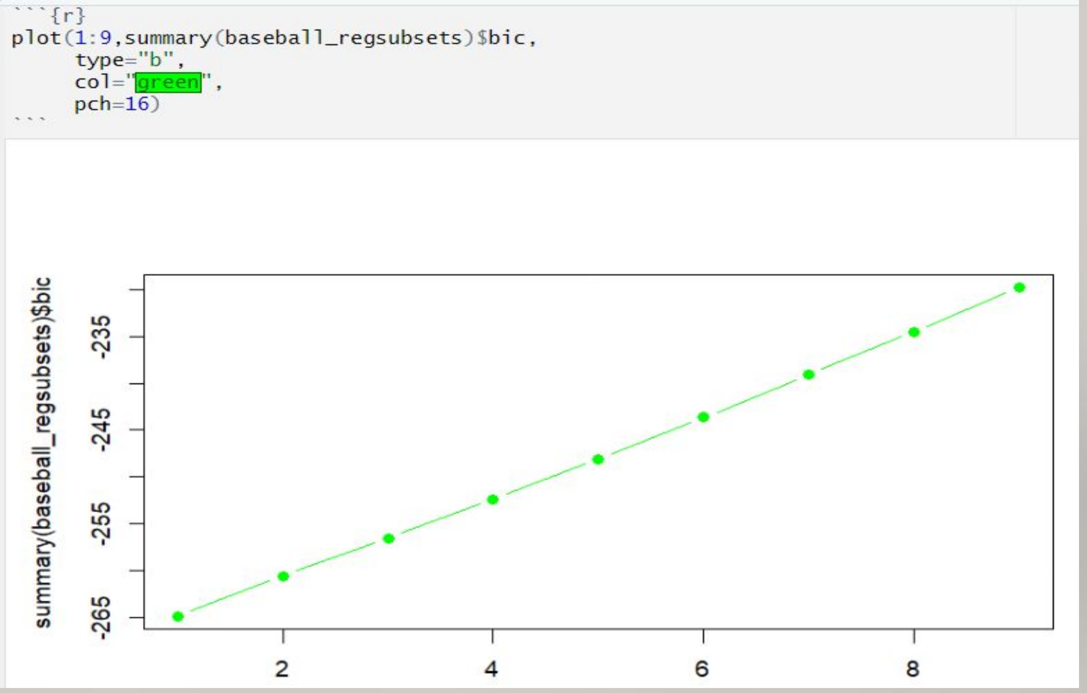
baseball_2:
Predicted WLR = -.1082+.76741(RunDiff)-.159(RunDiff)^2

Creating Models

Regsubset







Looking at the R^2 values, the model indicates it would be best to use all 9 predictor variables.

Looking at the Adjusted R^2 values, the model indicates it would be best to use only I predictor variable, which is RunDiff. .

Looking at the BIC values, the model indicates it would be best to use only I predictor variable, which is RunDiff.

Creating Models

Model Consensus

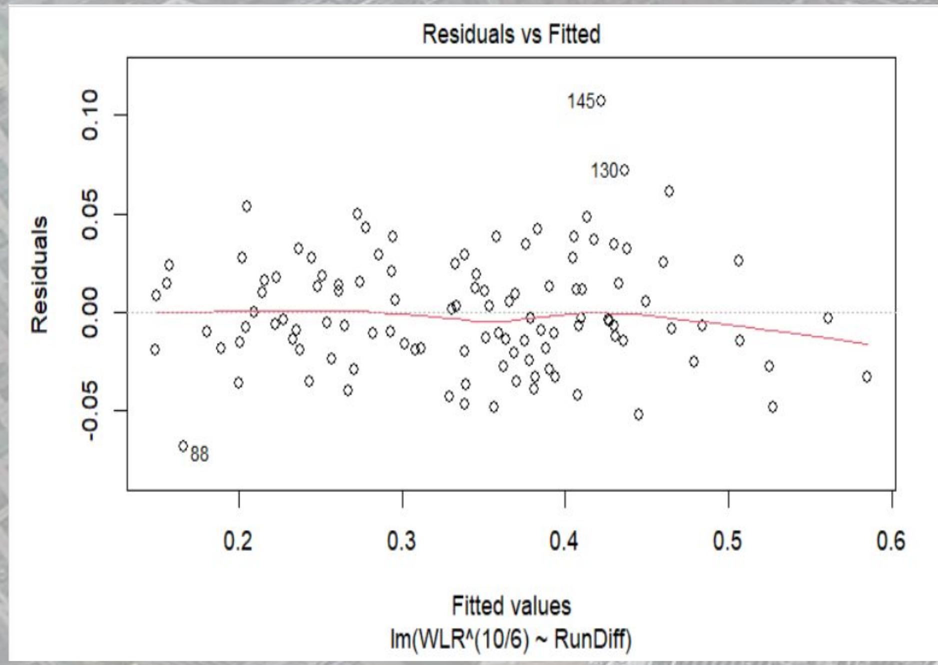
We have 3 models to consider using this point forward: baseball_1, baseball_2, and baseball_3

Baseball_3 is the most complex and has the lowest Adjusted R^2, so we will not use it.

Baseball_2 is slightly more complex than baseball_1, and has a slightly better Adjusted R^2. It isn't clear which of these two models (baseball_1 and baseball_2) may be better, so we will continue using both for the time being.

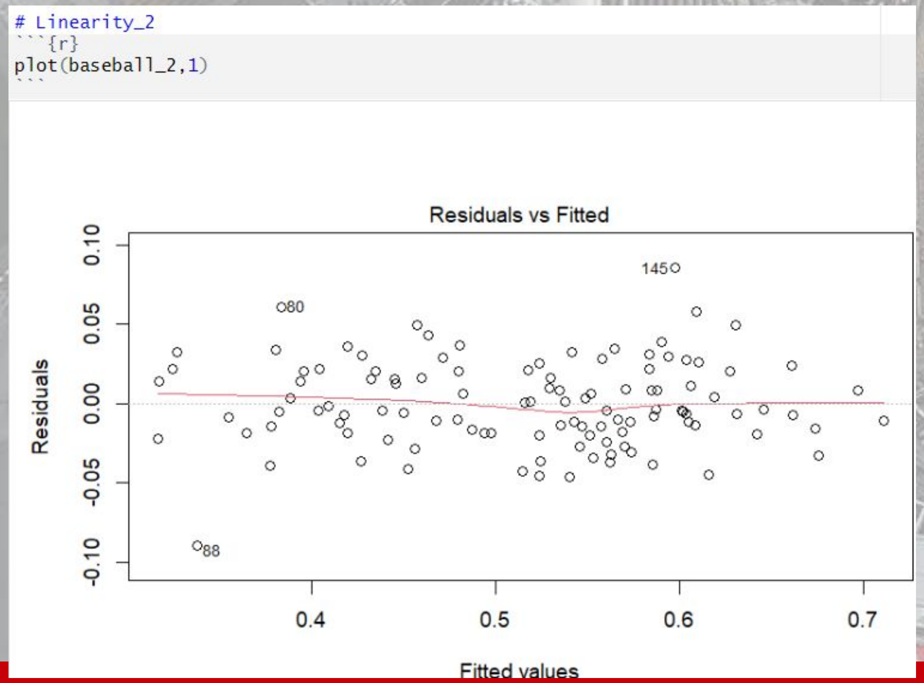
Model Conditions

Baseball 1



baseball_1 had violated the linearity condition, so it was refit using box-cox, then met linearity and constant variance. It had a high shapiro-wilk (.13) and Durbin Watson (.3)p-value, satisfying the normality and independence conditions. 91.91% of the variation in WLR is explained in baseball_1.

Baseball 2



baseball_2 satisfied the linearity and constant variance conditions. It also had a high shapiro-wilk (.4)and Durbin Watson (.32) p-value, satisfying the normality and independence conditions. 92.16% of the variation in WLR is explained in baseball_2.

Both Models were tested and found to be significant.

Mean Square Error

Comparing the mean square error between both models calculated against the 25% of testing data (0.0087 for baseball_2 vs 0.1712 for baseball_1), we concluded that model two (baseball_2) is the best model to fit our data to in order to predict Win-Loss Rate for the Atlanta Braves.

$$E(y) = -.1082 + .7674(RunDiff) - .159(RunDiff)^2$$

Conclusions