Stat 139 Final Project: EDA and Baseline Model

Danny Kim, Christopher Lee, Daniel Son, Karina Wang

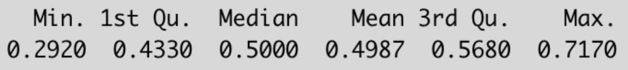
Description of Data and Source:

We pulled data in the form of csv files from https://www.baseball-reference.com/. For each year T, we use individual players' statistics from year T-1 in order to predict the team's performance in year T. Data from T = 2018, 2019, 2020, and 2021 were used for the training set, and data from T = 2022 were used for the test set. All exploratory data analysis was performed on the training set.

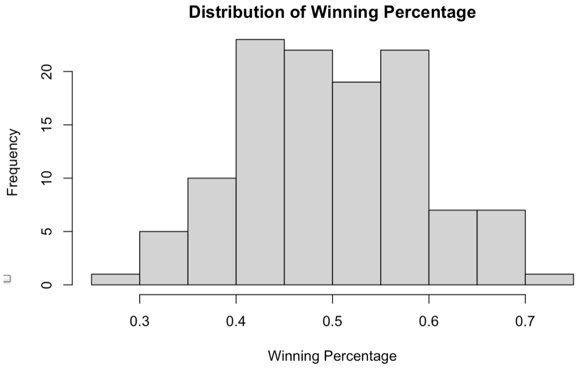
The source data listed each player's statistics individually and categorized players into either batters or pitchers. Thus, for each team, we averaged the individual players' data and then combined the batting and pitching statistics so that our data set had one row for every combination of team and year. We also dropped rows that didn't correspond to a team, for example rows that showed average league statistics.

Exploratory Data Analysis:

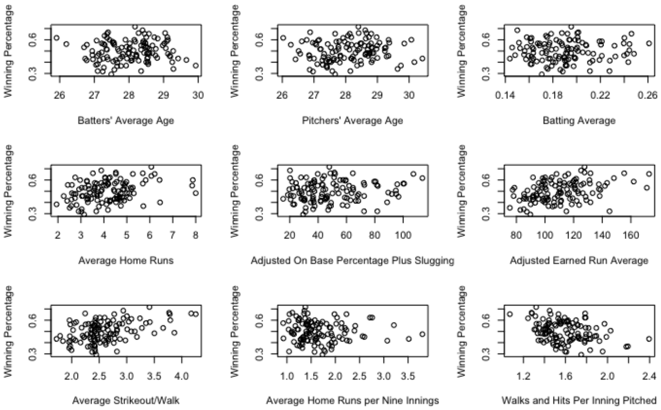
First, we explored the response variable of interest, which is the winning percentage of a team.



```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.2920  0.4330  0.5000  0.4987  0.5680  0.7170 
```



The distribution of winning percentage in the histogram above shows that it is symmetric about 0.5 as expected and roughly normal without any clear outliers, so we do not anticipate needing any transformations to this response variable. Next, we explored the relationship of several potential predictors with the response variable.

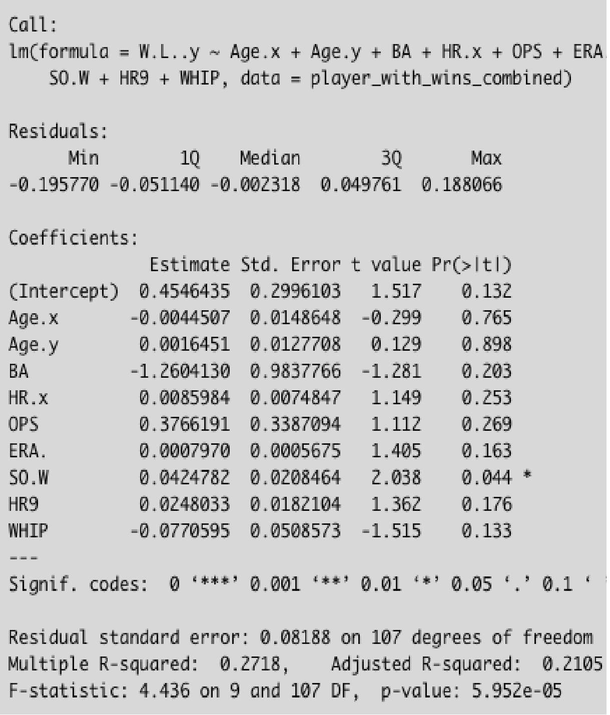


While none of the plots of the response versus predictors are clearly linear, the signs of the associations between the data make sense intuitively. For example, batters' average home runs and pitchers strikeout-to-walk ratio both seem to have a positive association with winning percentage. These two predictors are measures of the team's expected offensive and defensive strength, respectively, and thus we would expect these associations to be positive. Note that the signs of these correlations are not trivial since we are using the players' statistics from the previous year to predict the winning percentage of the current season; the fact that these associations are reasonable gives us reassurance that linear regressions will likely be a useful tool for a team's future performance.

Some predictors, namely HR.x (batters' home runs), OPS+ (adjusted on base percentage plus slugging percentage), and HR9 (pitchers' home runs allowed per nine innings) have far right outliers which we will take into consideration later on through transformation or censorship. For our baseline model to which we will compare all future models, we use the untransformed data.

Baseline Model:

For our baseline model, we fit a multiple linear regression model on the 9 predictors we explored in the Exploratory Data Analysis.

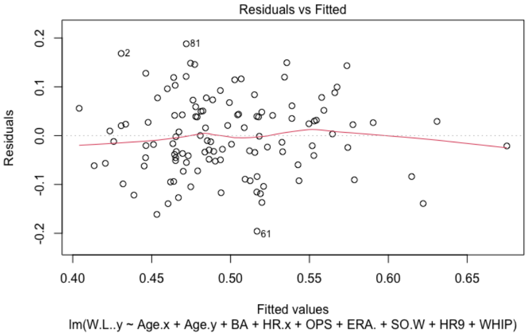


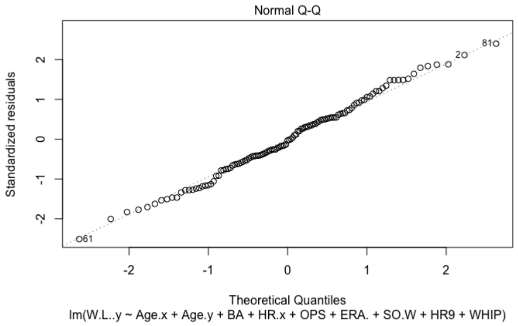
```
Call:
lm(formula = W.L..y ~ Age.x + Age.y + BA + HR.x + OPS + ERA.
    SO.W + HR9 + WHIP, data = player_with_wins_combined)

Residuals:
      Min        1Q    Median        3Q       Max
-0.195770 -0.051140 -0.002318  0.049761  0.188066

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4546435  0.2996103   1.517    0.132
Age.x       -0.0044507  0.0148648  -0.299    0.765
Age.y        0.0016451  0.0127708   0.129    0.898
BA          -1.2604130  0.9837766  -1.281    0.203
HR.x         0.0085984  0.0074847   1.149    0.253
OPS          0.3766191  0.3387094   1.112    0.269
ERA.         0.0007970  0.0005675   1.405    0.163
SO.W         0.0424782  0.0208464   2.038    0.044 *
HR9          0.0248033  0.0182104   1.362    0.176
WHIP        -0.0770595  0.0508573  -1.515    0.133
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 0.08188 on 107 degrees of freedom
Multiple R-squared:  0.2718,    Adjusted R-squared:  0.2105
F-statistic: 4.436 on 9 and 107 DF,  p-value: 5.952e-05
```

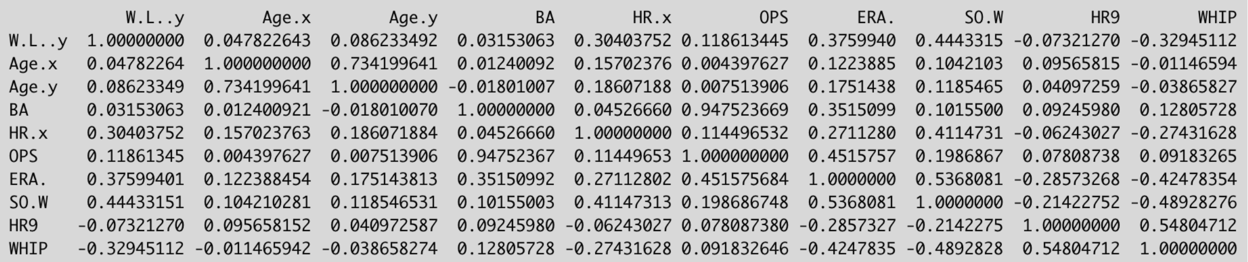




The nine predictors included in the multiple regression model are batter age, pitcher age, batting average (batters), home runs (batters), OPS+ (batters), adjusted earned run average (pitchers), strikeouts over walks (pitchers), HR9, and walks and hits per inning pitched (pitchers). These predictors are a mix of offensive and defensive statistics to encompass each of the important portions of the game. While there are many more options, we chose this set of predictors both because they seem intuitively important and because they each measure a different aspect of a team's players' skill sets.

In the model summary, we see that only the coefficient estimate on SO.W is statistically significant at the α = 0.05 level, with an estimate of 0.042, indicating that an increase of 1 in the pitchers' ratio of strikeouts to walks is expected to increase winning percentage by 0.042. This also means that in our current model, the direct effects of each of the other predictors is statistically insignificant, and thus we will investigate other possible predictors as well as feature engineering to gain better insight into the relationships between the data points. This baseline model has an R-squared of 0.272 and an adjusted R-squared of 0.211; the unadjusted value means that our linear model predicts 27.2% of the variance of the response variable. Since we do not anticipate transforming the response variable, we will compare all future models' performances to these values.

From the plots, we see that normality is roughly satisfied, though the residual plot shows some evidence of nonlinearity influenced by a few high outliers. We will consider this when building future models through data transformations. Constant variance is also roughly satisfied, with again some high outliers that deserve further future consideration. The data are not quite independent (an individual team is often similar between consecutive years), so we will use care when interpreting significance tests and confidence intervals. However, this independence violation will not bias our estimates, so we will still be able to make meaningful predictions from our models.



```
            W.L..y        Age.x        Age.y          BA        HR.x         OPS       ERA.       SO.W         HR9        WHIP
W.L..y  1.00000000  0.047822643  0.086233492  0.03153063  0.30403752 0.118613445  0.3759940  0.4443315 -0.07321270 -0.32945112
Age.x   0.04782264  1.000000000  0.734199641  0.01240092  0.15702376 0.004397627  0.1223885  0.1042103  0.09565815 -0.01146594
Age.y   0.08623349  0.734199641  1.000000000 -0.01801007  0.18607188 0.007513906  0.1751438  0.1185465  0.04097259 -0.03865827
BA      0.03153063  0.012400921 -0.018010070  1.00000000  0.04526660 0.947523669  0.3515099  0.1015500  0.09245980  0.12805728
HR.x    0.30403752  0.157023763  0.186071884  0.04526660  1.00000000 0.114496532  0.2711280  0.4114731 -0.06243027 -0.27431628
OPS     0.11861345  0.004397627  0.007513906  0.94752367  0.11449653 1.000000000  0.4515757  0.1986867  0.07808738  0.09183265
ERA.    0.37599401  0.122388454  0.175143813  0.35150992  0.27112802 0.451575684  1.0000000  0.5368081 -0.28573268 -0.42478354
SO.W    0.44433151  0.104210281  0.118546531  0.10155003  0.41147313 0.198686748  0.5368081  1.0000000 -0.21422752 -0.48928276
HR9    -0.07321270  0.095658152  0.040972587  0.09245980 -0.06243027 0.078087380 -0.2857327 -0.2142275  1.00000000  0.54804712
WHIP   -0.32945112 -0.011465942 -0.038658274  0.12805728 -0.27431628 0.091832646 -0.4247835 -0.4892828  0.54804712  1.00000000
```

When looking at the predictors, we find that the correlations between them are generally fairly low, with most being below 0.5 in magnitude. In particular, the correlations between the predictors within a predictor group (batters' stats or pitchers' stats) are largely similar in magnitude to the correlations between the predictors from different predictor groups. We expect that the offensive and defensive statistics should be independent of each other when conditioned on factors that affect the whole team, such as payroll. Since the within-group and between-group correlations are comparable in magnitude, this gives us some assurance that these predictors were chosen reasonably. In future analysis, we will look into the effect of other total team statistics and examine whether controlling for certain other factors reveals any insights into the association between the response variable and our current predictor set.

There is one pair of predictors, BA and OPS, which do have a high correlation (0.948), which makes sense given the similarities between the formulas used to calculate them. For future models, we will investigate whether we should continue to use both of them, only use one, or replace them with more fundamental measures of a team's ability to get on base and their power.