Predicting MLB Wins, Linear Regression

2024-04-21

Data Source = Lahman Baseball Package - Teams Data | Observational Unit: Baseball Team Data for seasons past 2000

Relevant Variables: yearID: year

franchID: franchise name abbreviated

divID: Division ID (either west, central, east) G: Games

Ghome: Games played at home W: Wins

L: Loses

DivWin: Categorical, binary yes or no whether team won their division

WCWin: Categorical, binary yes or no whether team won the Wild Card Game (postseason)

LgWIn: Categorical, binary yes or no whether team won their league (National or American League)

WSWin: Categorical, binary yes or no whether team won the World Series

R: A player is awarded a run if he crosses the plate to score his team a run.

AB: At bats

H: A hit (single, double, triple, home run)

X2B: A double

X3B: A triple

HR: Home run

BB: Walk

SO: Strike out

SB: Stolen bases

CS: Caught stealing

HBP: Hit by pitch

SF: Sacrifice Fly

RA: Runs allowed by team pitching

ER: Any run that scores against a pitcher without the benefit of an error or a passed ball.

ERA: Earned run average represents the number of earned runs a pitcher allows per nine innings with earned runs being any runs that scored without the aid of an error or a passed ball. ERA is the most commonly accepted statistical tool for evaluating pitchers.

CG: Complete game, the act of a pitcher pitching an entire game without the benefit of a relief pitcher. The total number of complete games thrown by a team's pitchers during the season.

SHO: A starting pitcher is credited with a shutout when he pitches the entire game for a team and does not allow the opposition to score.

SV: A save is awarded to the relief pitcher who finishes a game for the winning team. IPouts: Outs Pitched (innings pitched x 3)

HA: Hits allowed by pitchers

HRA: Home Runs allowed by pitchers

BBA: Walks allowed by pitchers

SOA: Strikeouts by pitchers

E: Errors

DP: Double Plays

Park: Team stadium

Attendance: Total attendance for team

BPF: Three-year park factor for batters

PPF: Pitching Park Factor - centered around 100, with numbers above 100 representing the percentage increase in run-scoring against pitchers in that park as compared to other parks, and numbers below 100 representing the percentage decrease in run-scoring against pitchers in that park. A metric for how hitter or pitcher-friendly a team's park is.

RD = Run Differential: Runs scored by batters minus runs allowed by pitchers

wpct= Win Percentage (runs divided by the sum of runs scored and runs allowed)

expwin = Pythagorean expectation is a formula invented by Bill James to estimate how many games a baseball team "should" have won based on the number of runs they scored (R) and allowed (RA).

diff: Wins minus expected wins. Comparing a team's actual and Pythagorean winning percentage can be used to evaluate how lucky that team was (by examining the relation between the two winning percentages).

BA = Batting average, number of hits divided by number of at bats. The most commonly used batting statistic.

OBP = On base percentage. Sum of all on base scenarios (notably including walks) divided by the total number of at bats.

SLG = An evaluative metric to determine player power. Singles, doubles, home runs, etc are weighted differently.

OPS = On base percentage + slugging. An statistic to easily evaluate overall offensive performance wOBA = Weighted on base average. It is formed from taking the observed run values of various offensive events, dividing by a player's plate appearances, and scaling the result to be on the same scale as on-base percentage.

ISO = Isolated power. Measures how many extra bases a player averages per at bat. BABIP = Batting average on balls in play. Used to determine how "lucky" a batter is. High BABIP = more lucky

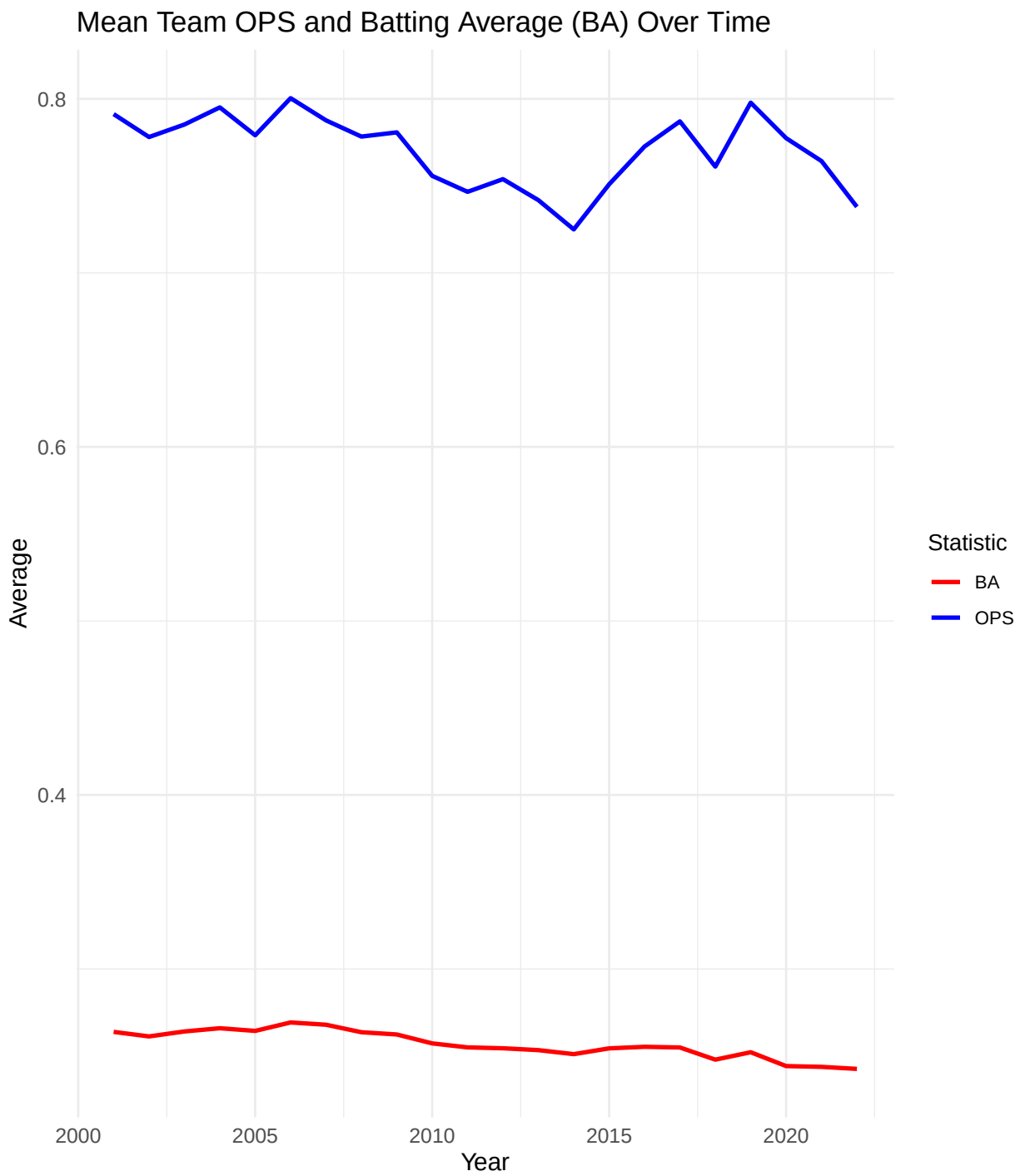
FIP = Fielding independent pitching. Attempts to use the ERA scale to more accurately reflect events under the pitcher's control: home runs, strikeouts, walks (intentional ones stripped out) and hit-by-pitch.

WHIP = Walks plus hits per inning pitched (WHIP) is a measurement of the number of baserunners a pitcher has allowed per inning pitched.

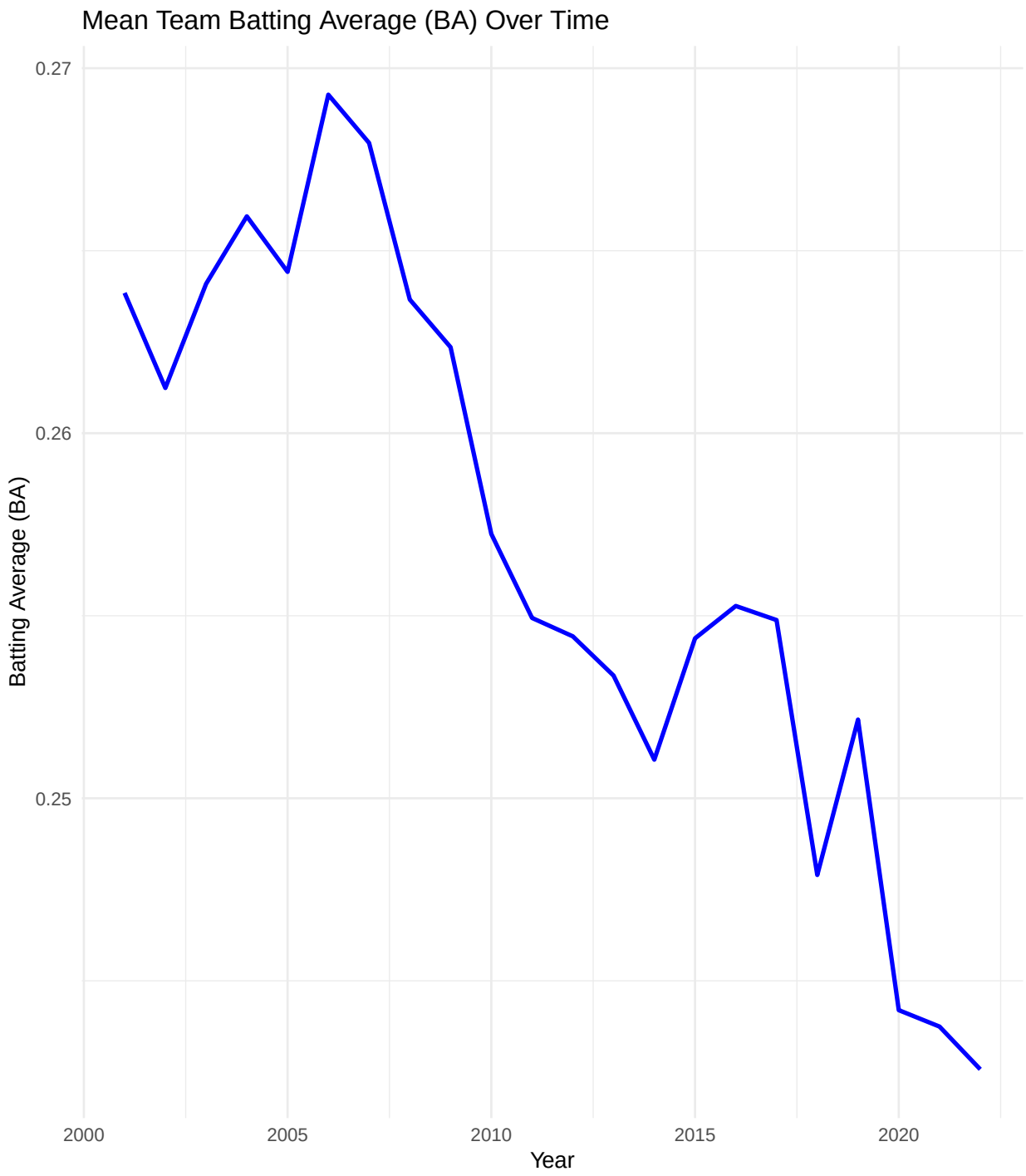
Ks_pitcher = Strikeout rate for pitchers

K_rate = Strikeout rate for batters

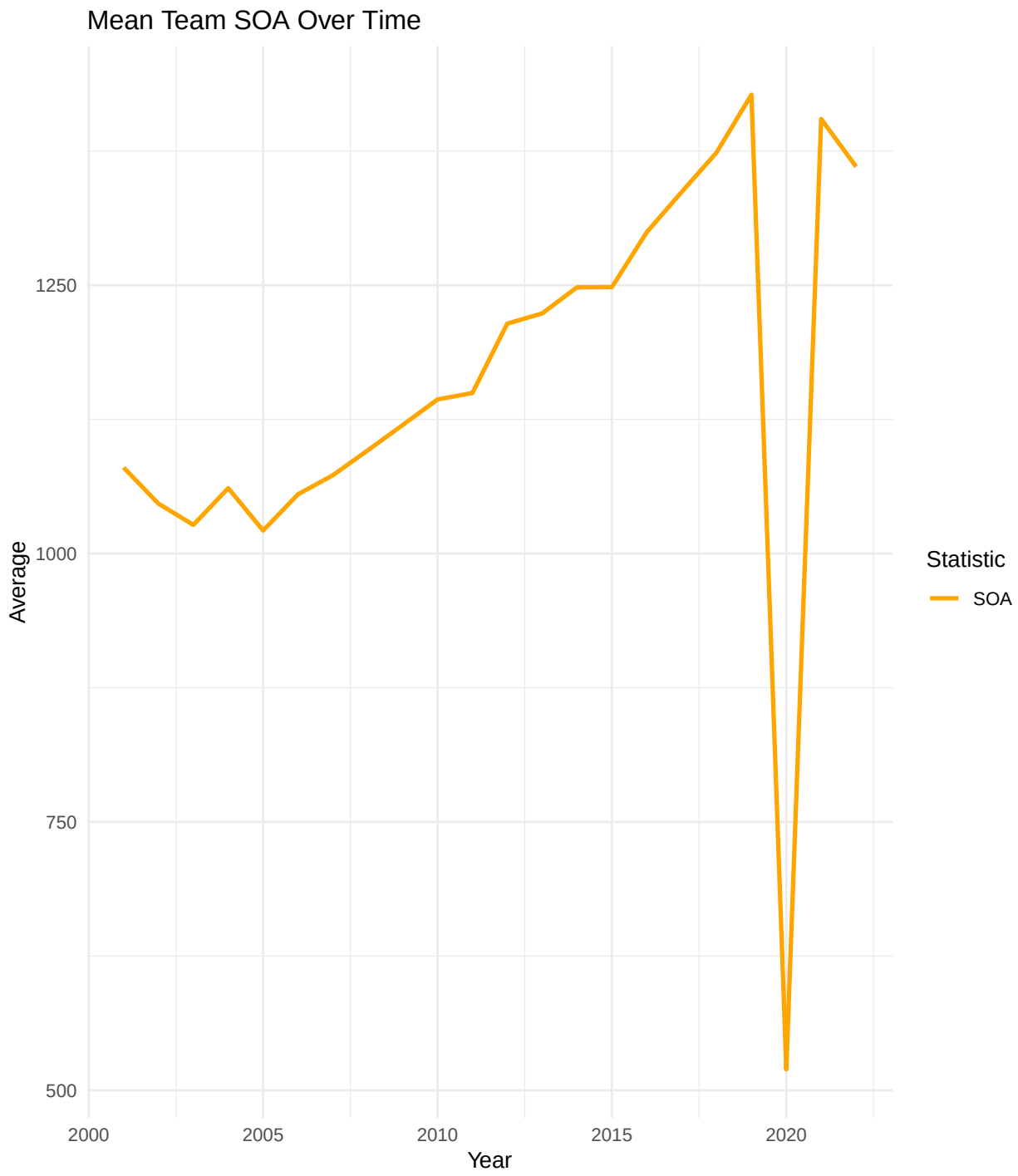
Attendance_Quality = Takes the quratiles of fan attendance and classifies as either "poor", "average", or "good".



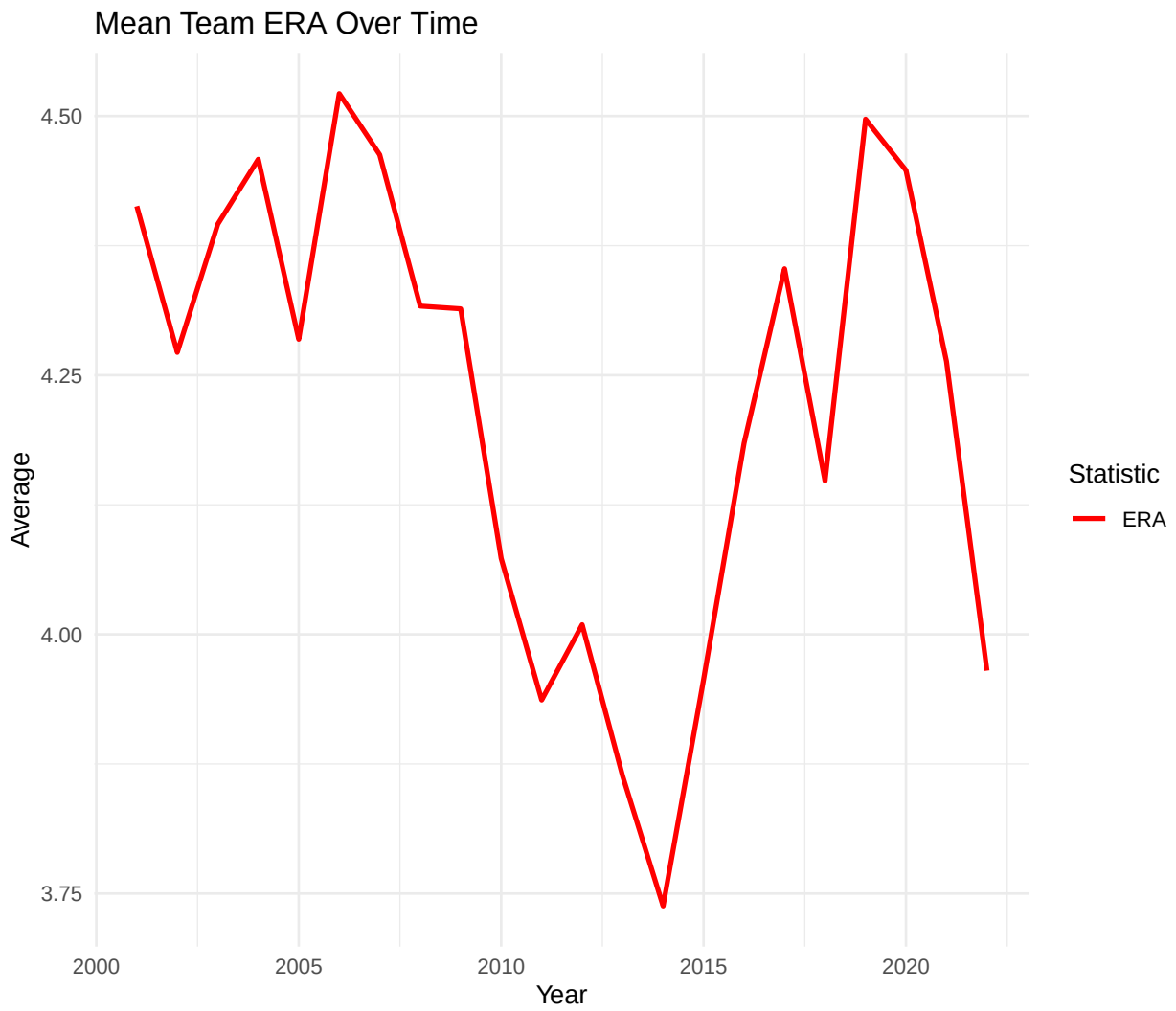
Team hitting quality is seeming to decrease. Wins nowadays could be less dependent on offensive rather than pitching



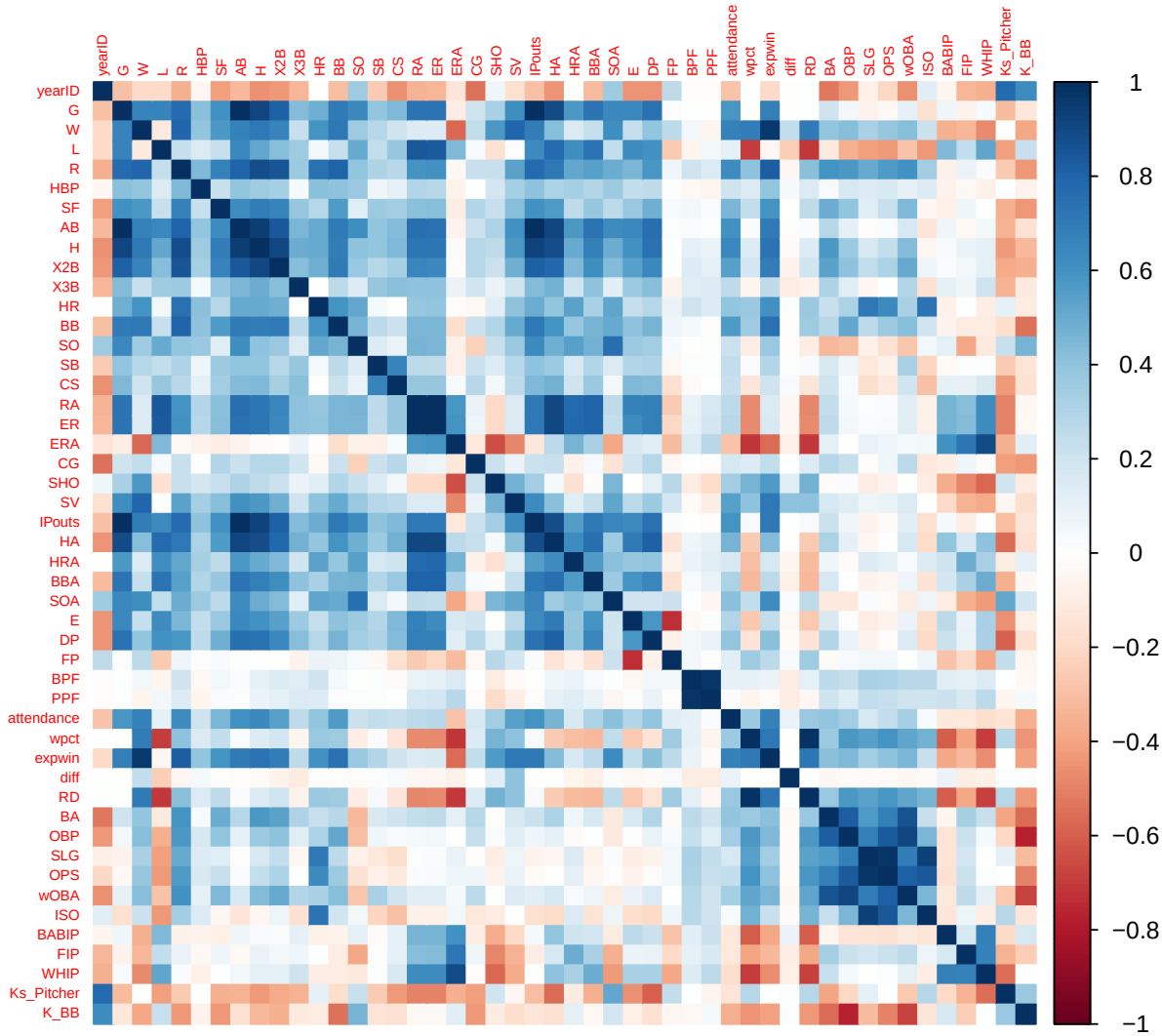
Batting quality has decreased over time.



Team strikeouts have increased over time (excpet for 2020, pandemic season)



Strikeouts (except for the year 2020 where there was a 60 game season) are increasing, and ERA is decreasing over time meaning pitch quality is increasing.



Correlation matrix: RD Wpct correlated with ERA. K/BB rate is correlated with OBP. WPCT correlated with wins.

As a baseball fan, none of these results particularly surprised me. It is a known fact that pitching quality has increased over time, subsequently impacting hitting quality. I am curious about the 2020 season though, which was impacted by COVID and led to a 60 game season. Were there certain predictors that impacted wins during a 60 game season more than a 162 game season? I think I got a very representative sample of my population, not just because the data are team totals for these metrics, but also since the data is from 2001 onward. The sheer quantity of baseball data makes it a viable source to extract meaningful insights.

###Part 2:

High OBP correlated with winning, seems like FIP and OBP are correlated which seems wrong... Wins and expected wins are correlated, which is a good sign because it shows that Runs and Runs Allowed are good indicators of team success. ggpairs(selected_vars5) shows data clumped at the bottom.

From the original "Teams" dataset, I derived some commonly used baseball statistics which I defined earlier. Expwin, diff, RD, BA, OBP, SLG, OPS, wOBA, ISO, BABIP, FIP, WHIP, Ks_Pitcher, K_BB, and Attendance_Quality. These are interaction variables because they exist and were created in the context of the original counting statistics.

I think these variables are necessary because they provided context to the original counting stats. For example, if there are more team hits, team BA (batting average) would increase.

```
![](Project---Win-Predictions_files/figure-latex/figs-1.pdf)<!-- -->

```
```
##
## Call:
## lm(formula = W ~ SV + R + E + SHO + RA + IPouts + WHIP + BABIP +
##     HRA + HA + Ks_Pitcher + BA + BBA, data = data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4835 -1.9498 -0.1505  1.9647  8.8504
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.186859  12.088802  -0.595  0.55249
## SV            0.400394   0.024893  16.085  < 2e-16 ***
## R             0.090885   0.002935  30.970  < 2e-16 ***
## E            -0.024649   0.009980  -2.470  0.01391 *
## SHO           0.146253   0.047331   3.090  0.00213 **
## RA           -0.042636   0.006291  -6.777 4.10e-11 ***
## IPouts        0.016987   0.002935   5.787 1.39e-08 ***
## WHIP        -48.265240  11.421152  -4.226 2.91e-05 ***
## BABIP       274.955458  45.001509   6.110 2.25e-09 ***
## HRA           0.018829   0.011148   1.689  0.09193 .
## HA           -0.045035   0.009141  -4.927 1.20e-06 ***
## Ks_Pitcher  -26.163710   4.802560  -5.448 8.64e-08 ***
## BA           74.323355  23.218418   3.201  0.00147 **
## BBA           0.019900   0.008565   2.324  0.02062 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.918 on 426 degrees of freedom
## Multiple R-squared:  0.9663, Adjusted R-squared:  0.9653
## F-statistic: 940.6 on 13 and 426 DF,  p-value: < 2.2e-16
```

I split my modified dataset into two separate datasets to separate any interaction variables.
Then, used best subset selection for variable selection sorted by lowest BIC.

After, I took the variables selected by each best subset (regsubset) and combined them into one model,
and then used forward-backward selection to select the most significant variables from there.

Residuals are randomly distributed about y=0.

The most significant variables are R, RA, IPouts, WHIP,

BABIP, HA, Ks_Pitcher. But Errors, SHO, BA, and BBA are also significant, just less so.
Most of these metrics are pitching based, which tracks
with my assessment earlier that pitching quality is increasing.
It would make sense that certain hitting metrics are significant indicators of team wins.
But, Ks_Pitcher is a strange instance because typically, you would expect more strikeouts
to correlate with more wins.
So I will choose to omit it from the model.

Additionally, in the mid-1990's, a man named Bill James
developed a formula that predicted the percentage of games a team is to win based on the
number of runs scored and runs allowed:
(runs scored)^2 = (runs scored)^2 +(runs allowed)^2, where runs allowed =0.
My model includes both Runs Scored and Runs Allowed as
significant indicators of team wins, which is on par with Bill James' famous formula.

My final model: W ~ SV + R + RA + IPouts + WHIP + BABIP + HA + Ks_Pitcher

```
##
## Call:
## lm(formula = W ~ SV + R + RA + IPouts + WHIP + BABIP + HA + Ks_Pitcher,
##     data = data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2815 -1.9095  0.0087  1.9879  9.2772
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.378814   6.919656  -0.922    0.357
## SV            0.406707   0.025257  16.103  < 2e-16 ***
## R             0.096065   0.002248  42.726  < 2e-16 ***
## RA           -0.053549   0.005165 -10.367  < 2e-16 ***
## IPouts        0.018738   0.001777  10.542  < 2e-16 ***
## WHIP        -18.970424   4.410904  -4.301 2.11e-05 ***
## BABIP       161.370320  28.913326   5.581 4.23e-08 ***
## HA           -0.036403   0.007456  -4.882 1.48e-06 ***
## Ks_Pitcher  -14.860574   3.448481  -4.309 2.03e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3 on 431 degrees of freedom
## Multiple R-squared:  0.964,  Adjusted R-squared:  0.9633
## F-statistic:  1442 on 8 and 431 DF,  p-value: < 2.2e-16
```

Interpreting the Beta coefficients:

SV (Saves): For each additional save, the number of wins is expected to increase
by approximately
0.419 holding all other variables constant.
This is highly significant, with a p-value < 0.001

showing a strong positive relationship between saves and wins.

R (Runs): Each additional run is associated with an increase of about 0.09 wins.
This is also highly significant with a p-value < 0.001.

RA (Runs Allowed by Pitchers): This coefficient is negative, meaning that
as runs
allowed increase,
the number of wins decreases.
Each run allowed by pitchers is associated with -0.068 wins. The p-value < 0.01,
and is highly significant.

IPouts (Innings Pitched Outs): More outs pitches is associated with more wins,
increasing approximately 0.013 wins for each out pitched, highly significant
p-value < 0.001

WHIP (Walks Plus Hits Per Inning Pitched): Significant negative relationship.
A high WHIP decreases wins significantly.
This indicates a significant negative relationship between WHIP and the number
of wins.
Specifically, for each unit increase in WHIP the number of wins
is expected to decrease by approximately -10.05 assuming all other variables are held constant.
The difference between a 0.00 WHIP, 1.00 WHIP, and 2.00 WHIP is massive.

BABIP (Batting Average on Balls in Play): High BABIP increases wins
significantly.
The scale of BABIP typically ranges from 0 to 1, as it represents the batting
average of a player excluding home runs and strikeouts.
Given this range, a coefficient as large as 56.810 suggests that even a small
change in
BABIP could result in a large change in the predicted number of wins.

HA (Hits Allowed by Pitchers): More hits allowed by pitchers, fewer wins.
Significant p value < 0.05. For each hit allowed by pitchers,
it is expected for wins to decrease by -0.008921, holding all other variables
constant.

My R-Squared is 0.9621022, which is high, which means my model explains a
significant portion of the variability in the number of wins
based on my predictors. While a high R^2 suggests a good fit to the data
I tested it on, it doesn't necessarily guarantee that the model will accurately
describe the population or perform well on new, unseen data. If a model is too
closely fitted to the training data, it might capture noise as well as the
actual significant predictors. This can lead to excellent performance on
training and test data
but poor performance on any new data.
But overall, a high R^2 on my test data is promising.

The residuals are clustered around higher predicted wins. This suggests that my model is making more predictions in this range, which makes sense because most team wins fall in this range. Additionally, the residuals on the right (higher wins) do not show any systematic patterns (like trends or increasing variance). This suggests that there isn't a bias or changing variance at this end of the scale, which is positive.

Standardized residuals provide a way to assess the relative size of the residuals in terms of standard deviations, making it easier to identify outliers. There only seems to be one outlier, but in the nature of baseball, outliers exist and should be accounted for.

```
##           SV            R           RA       IPouts         WHIP        BABIP
##   39.3113636  711.9522727  715.1840909 4210.4613636    1.3455847    0.2970823
##           HA   Ks_Pitcher           BA
## 1384.9318182    0.8263019    0.2568497
```

```
##        fit      lwr      upr
## 1 78.32273 78.04161 78.60384
```

```
##        fit      lwr      upr
## 1 78.32273 72.41929 84.22617
```

Mean Predicted Value: This interval suggests that the true mean response, given the predictors set at their mean values, is expected to lie within this range (~78-~84 wins) with a 95% confidence level. The narrowness of this interval indicates a high level of precision in estimating the mean response from the model.

Future Prediction Interval: This interval is wider than the confidence interval for the mean, reflecting not only the uncertainty in estimating the mean response but also the additional variability of individual future observations around this mean.

This interval tells us that if we were to observe a new data point with the predictors set at the same mean values, we would expect it to fall within this range (~78- ~ 84 wins) with a 95% level of confidence.

Interpretation of Model:

The variables in my final model:

SV (Saves): Directly impacts games won, as a save is recorded only when a pitcher finishes a game his team wins.

R (Runs): More runs scored typically lead to more games won.

RA (Runs Allowed): Fewer runs allowed by the defense usually means more wins.

IPouts (Innings Pitched Outs): Indicates the durability and effectiveness of the pitching staff, affecting game outcomes.

WHIP (Walks Plus Hits per Inning Pitched): Lower WHIP values suggest better pitching control and effectiveness, leading to fewer opponents scoring.

BABIP (Batting Average on Balls in Play): Reflects how often a team in play (excluding home runs) gets hits, influencing game outcomes.

HA (Hits Allowed): Fewer hits allowed generally leads to fewer runs scored by the opposition.

Variables not included in the final model might have been dropped due to:

Redundancy: Some variables might provide overlapping information. For example, ERA (Earned Run Average) and WHIP both measure pitching effectiveness but in slightly different ways. The model might favor one over the other if they are highly correlated.

Lack of Additional Predictive Power: Some variables, while potentially relevant, might not provide significant predictive power beyond what is already explained by the included variables.

Correlation Among Variables:

Potential High Correlation: Variables like ERA, WHIP, and FIP might be correlated since they all relate to pitching effectiveness. Similarly, OBP, BA, and SLG might be correlated as they relate to offensive performance.

Impact of Correlation: High correlation among variables can lead to multicollinearity, where it becomes difficult to isolate the effect of each variable. This might lead to some variables being dropped from the model if they do not uniquely contribute to explaining the variance

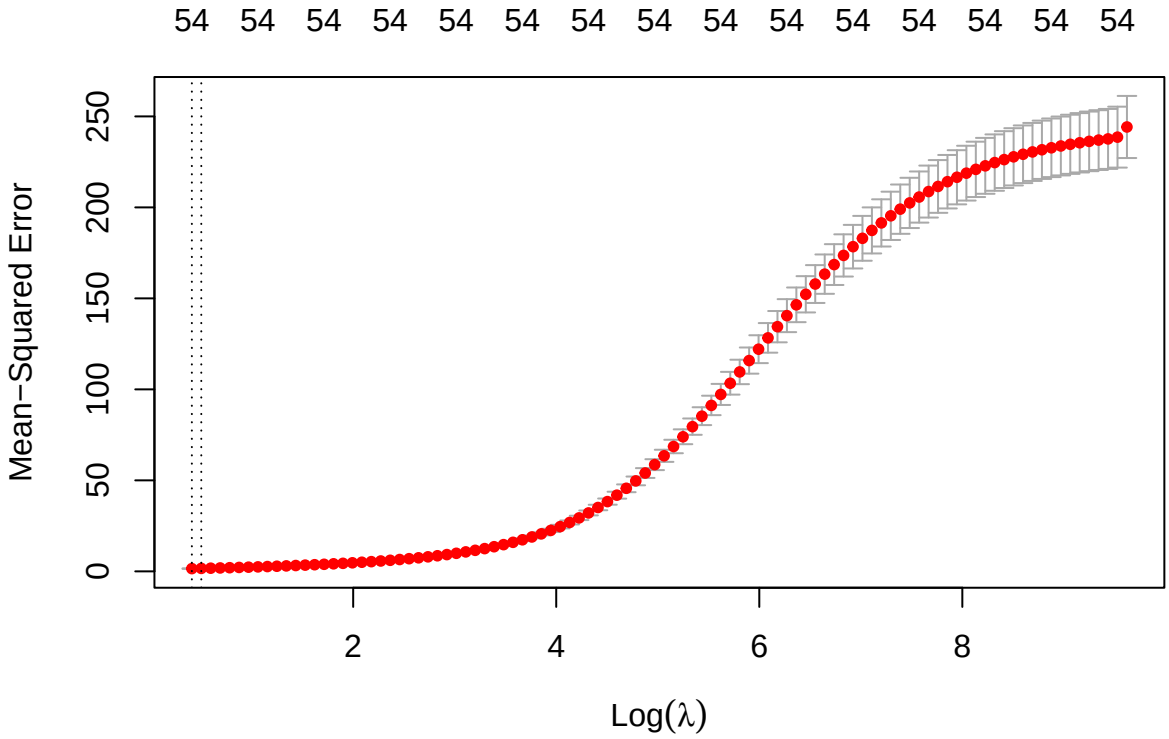
Part 3

final_model <- lm(formula = W ~ SV + R + RA + IPouts + WHIP + BABIP + HA + Ks_Pitcher, data = data_train)

Predicting Wins and these are the most significant predictors: -SV (Saves) -R (Runs) -RA (Runs Allowed by pitchers) -IPouts (Outs Pitched (innings pitched x 3) -WHIP (Walks and Hits per Innings Pitched) -BABIP (Batting average on balls in play) -HA (Hits allowed by pitchers) -Ks_pitcher (Strikeouts by pitcher)

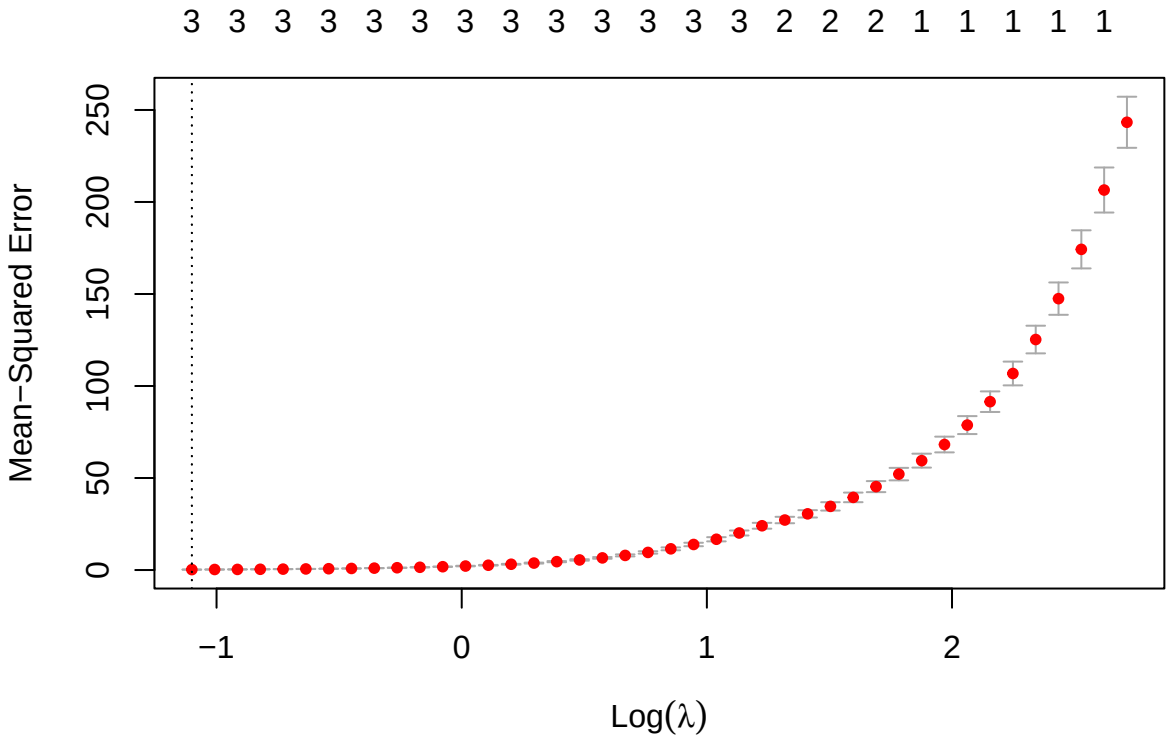
This data comes from the Lahman baseball package. I also did some feature engineering to make WHIP, BABIP, and Ks_Pitcher

I am only using data past the year 2000.




```
## [1] "Best lambda: 1.508311625254"
```

Lasso




```
##          s1
## 2  91.71626
## 4  63.57630
## 5  82.09556
## 6  82.91005
## 13 92.36681
## 20 81.41212
```

Compare models

MLR (final_model):

R-squared: 0.964 Adjusted R-squared: 0.9653 MSE: 10.4499670455261 RMSE: 3.23264087790867

Given the variance of W is approximately 255.17 and my MSE is 10.45, the ratio of MSE to the variance is about 0.041. This ratio is quite low, indicating that the Mean Squared Error of my model is small relative to the variance of the response variable This suggests that your model has significant predictive value and is performing well in terms of error magnitude relative to the natural variability in the data.

Coefficients: SV, R, RA, IPouts, WHIP, BABIP, HA, Ks_Pitcher

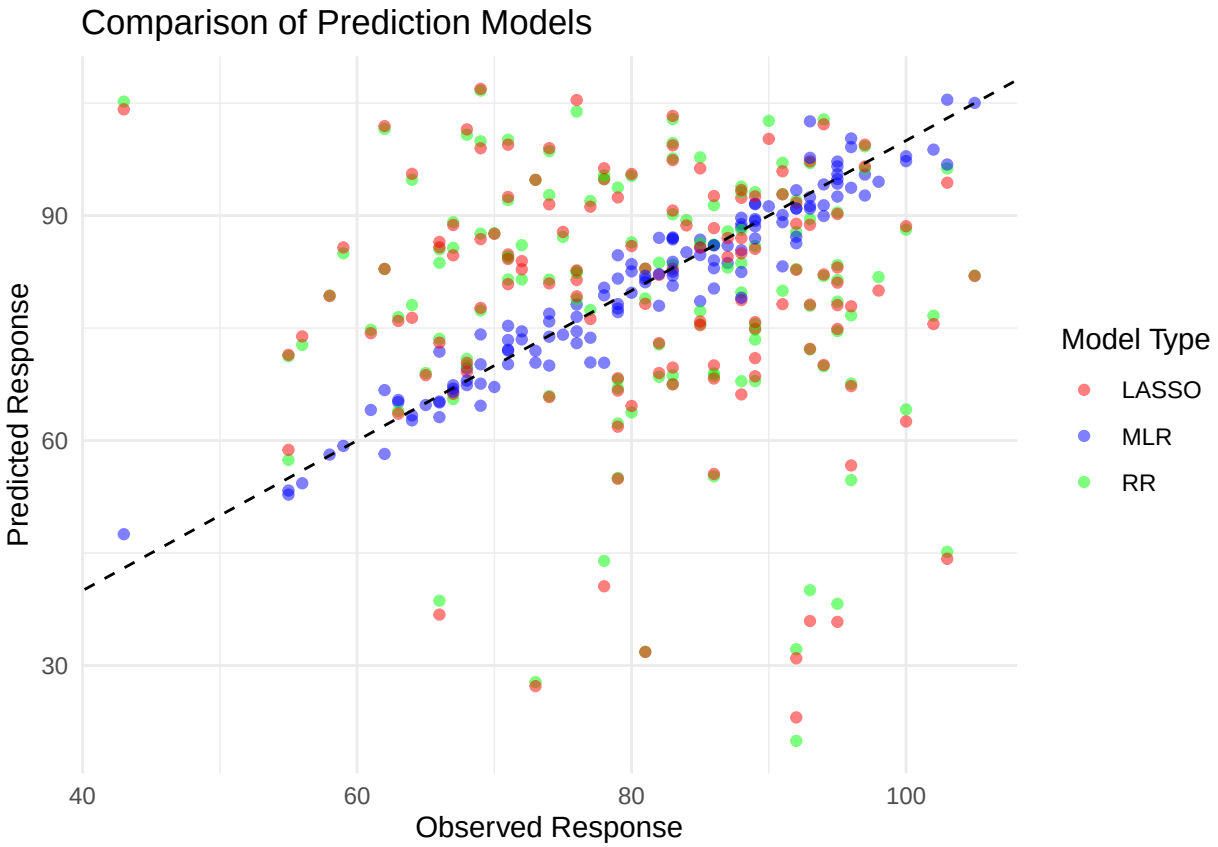
Ridge Regression:

R-squared: 0.973940413040255 Adjusted R-squared: 0.9262392343866 MSE: 1.70366541619586" RMSE: 1.19888363620898"

LASSO

R-squared: 0.999182054984301 Adjusted R-squared: 0.99761887117652 MSE: 0.198471103686519" RMSE: 0.445501

The coefficients selected by LASSO were: SV (Saves), expwin (Expected Wins), diff (Run Differential)



RR and LASSO Performance: Closeness to y=x. When the predictions from the RR and LASSO models are close to y=x, the models are performing well. y=x represents perfect predictions where the predicted values exactly match the observed values. Being close to this line indicates that the models have a high accuracy in terms of prediction. Both RR and LASSO incorporate regularization, which helps in reducing overfitting. This is effective because my dataset has multicollinearity and is high-dimensional. The regularization might be helping these models generalize better on the test data.

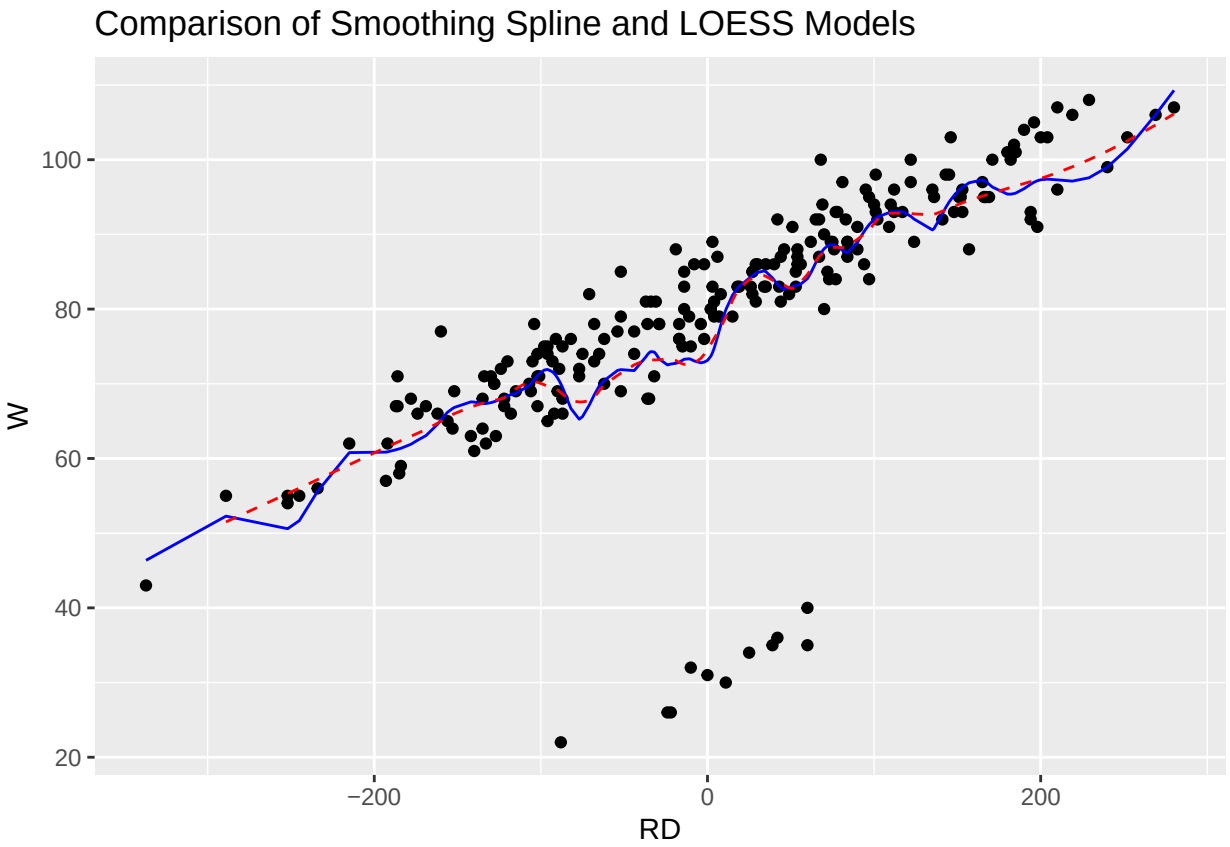
###MLR Performance: Scattered Plot Points: The MLR predictions are more scattered and deviate from the y=x line. This suggests that the MLR model might be experiencing overfitting or underfitting, or it might not be capturing all the relevant patterns in the data. This could be due to:

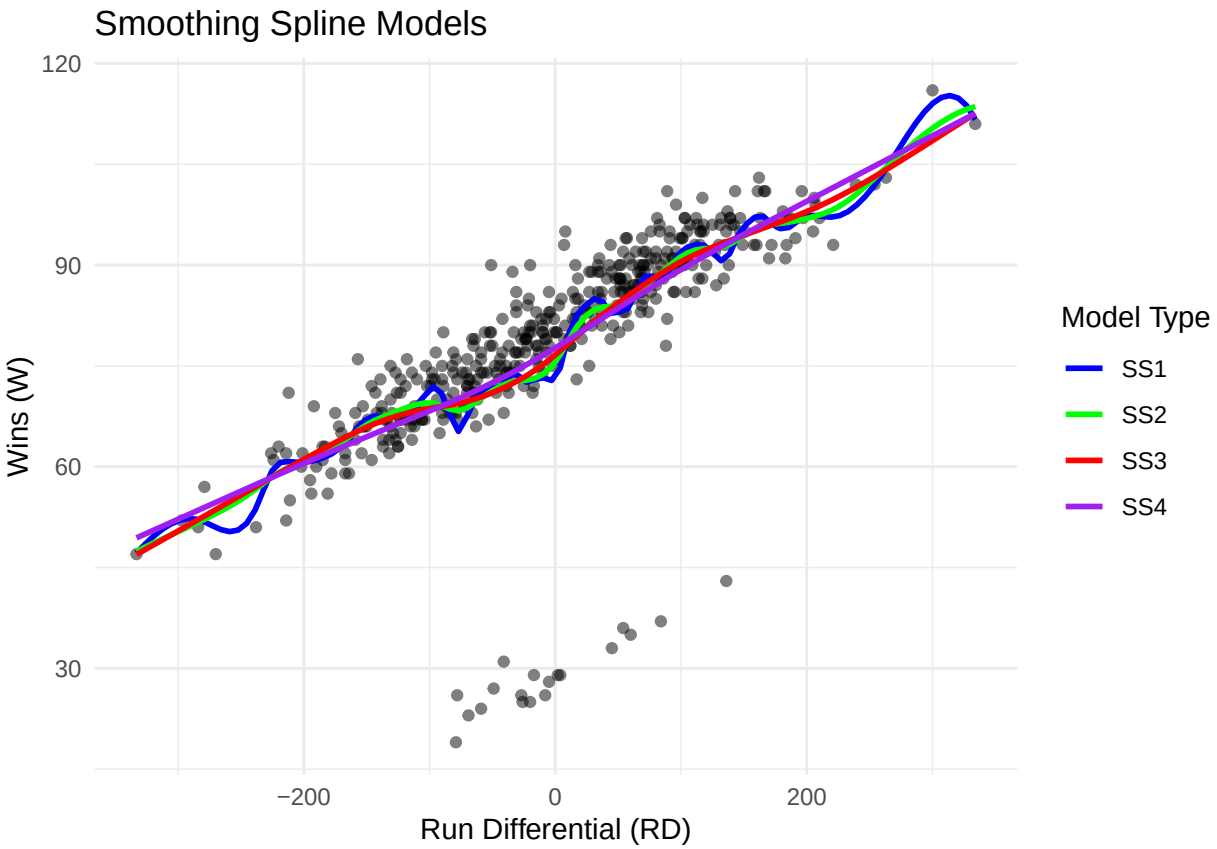
Lack of Regularization: Unlike RR and LASSO, standard MLR doesn't include regularization which can lead to overfitting, especially if there are many predictors or if the predictors are highly correlated. Model Complexity: MLR might be too simple or too complex for the data structure, failing to capture the essential relationships between variables.

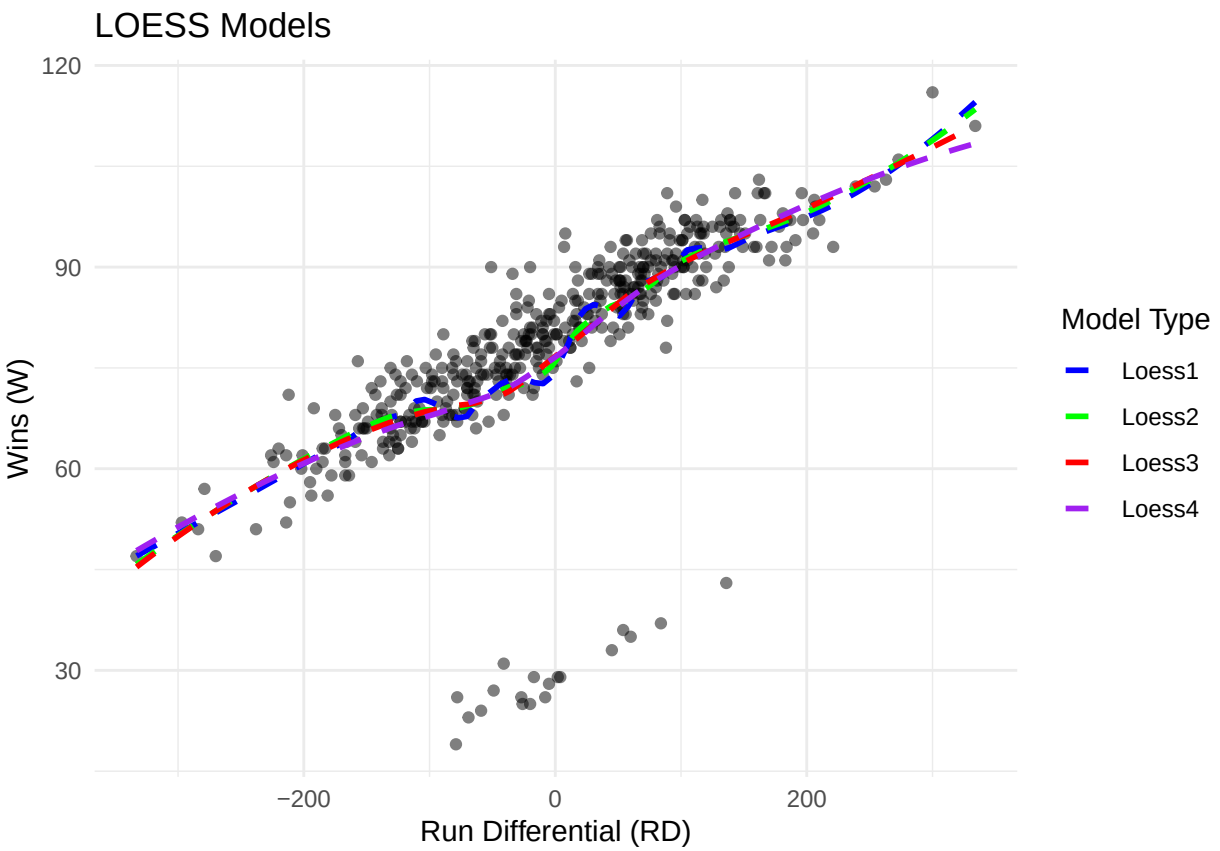
#Run both smoothing spline and loess smoother models

```
## Warning: Use of `data_test$RD` is discouraged.
## i Use `RD` instead.
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_line()`).
```


Comparison of Smoothing Spline and LOESS Models
100
80
60
40
20
W
−200
0
200
RD


Smoothing Spline Models
120
90
60
30
Wins (W)
−200
0
200
Run Differential (RD)
Model Type
SS1
SS2
SS3
SS4



Smoothing Spline Plot: Lower spar values may show more variability following the data closely, while higher values result in smoother curves. These models are generally good at providing a balance between smoothness and fit. The spar parameter controls this balance:

LOESS Plot: The span parameter variations will show similar trends where smaller spans fit the data more closely, capturing more fluctuations, and larger spans smooth out these fluctuations, providing a general trend. LOESS is particularly flexible and locally adaptive, making it excellent for datasets with varying trends across the domain.

Given the considerations of smoothness, interpretability, and the ability to capture variability, I would recommend choosing the LOESS model with a span of 0.6 (loess_model3) for future predictions. This model offers a balanced approach with moderate smoothing that captures significant trends without overly fitting minor fluctuations in the data. The flexibility of LOESS allows it to adapt well to the underlying patterns in the data, making it suitable for datasets with complex relationships that do not conform to a specific functional form. Additionally, the moderate span helps in maintaining a good balance between bias and variance, providing a reliable model for predicting future outcomes based on the run differential. This choice is based on the assumption that the model performs consistently across different data segments and aligns closely with observed values in visual assessments.

##Conclusion

I was suprised that the MLR performance wasn't as good as the model summary indicated it was. LASSO provided the best results for feature selection which led to a simplified, still effective model. Because my data was only from the year 2000 onwards, I am curious if during different time periods, different variables would significantly contribute to wins.

##Logistic regression: Will this team make the playoffs?

```
## geom_text_repel: parse = FALSE, na.rm = FALSE, box.padding = 0.25, point.padding = 1e-06, min.segmen
## stat_identity: na.rm = FALSE
```

```
## position_identity

## Error in `$<-.data.frame`(`*tmp*`, DivWin, value = logical(0)): replacement has 0 rows, data has 220

## Error in `$<-.data.frame`(`*tmp*`, DivWin, value = logical(0)): replacement has 0 rows, data has 440

## Error in eval(substitute(select), nl, parent.frame()): object 'DivWinY' not found

##
## Call:
## glm(formula = DivWin ~ SV + R + RA + IPouts + WHIP + BABIP +
##     HA + Ks_Pitcher + RD + diff, family = binomial(), data = data_train)
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.383235   8.134484  -0.785    0.433
## SV           -0.017730   0.038504  -0.460    0.645
## R             0.027668   0.004025   6.874 6.25e-12 ***
## RA           -0.048421   0.008839  -5.478 4.30e-08 ***
## IPouts       -0.001449   0.002183  -0.664    0.507
## WHIP          6.006483   5.712578   1.051    0.293
## BABIP       -37.508496  40.294420  -0.931    0.352
## HA            0.015599   0.010416   1.498    0.134
## Ks_Pitcher    7.076874   4.876263   1.451    0.147
## RD                 NA         NA      NA       NA
## diff          0.337184   0.066618   5.061 4.16e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 434.75  on 439  degrees of freedom
## Residual deviance: 210.17  on 430  degrees of freedom
## AIC: 230.17
##
## Number of Fisher Scoring iterations: 7
```

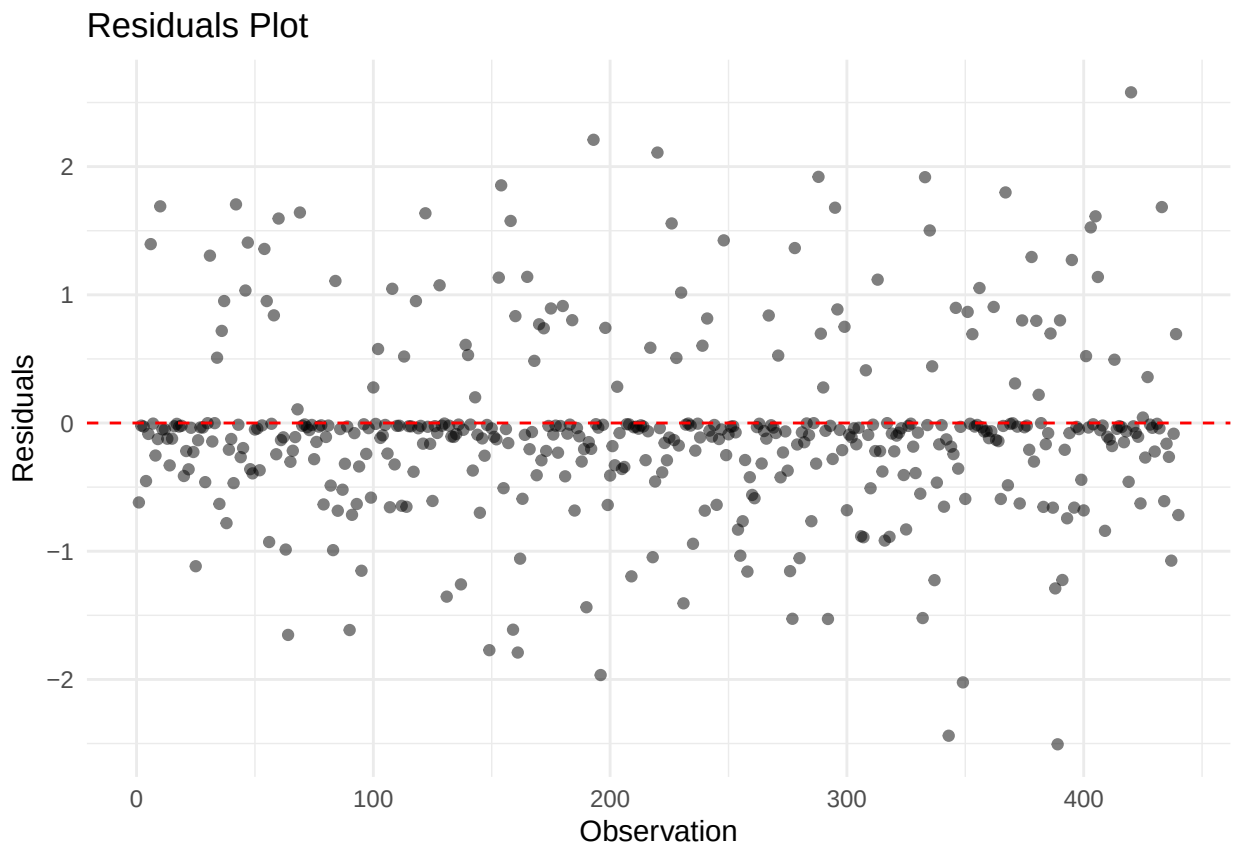
These coefficients represent the change in the log odds of the dependent variable DivWin for a one-unit change in the predictor variables, holding all other predictors constant.

For example, the coefficient for R (Runs scored) is 0.027668, indicating that for each additional run scored, the log odds of winning (DivWin) increase, suggesting a positive relationship between runs scored and winning.

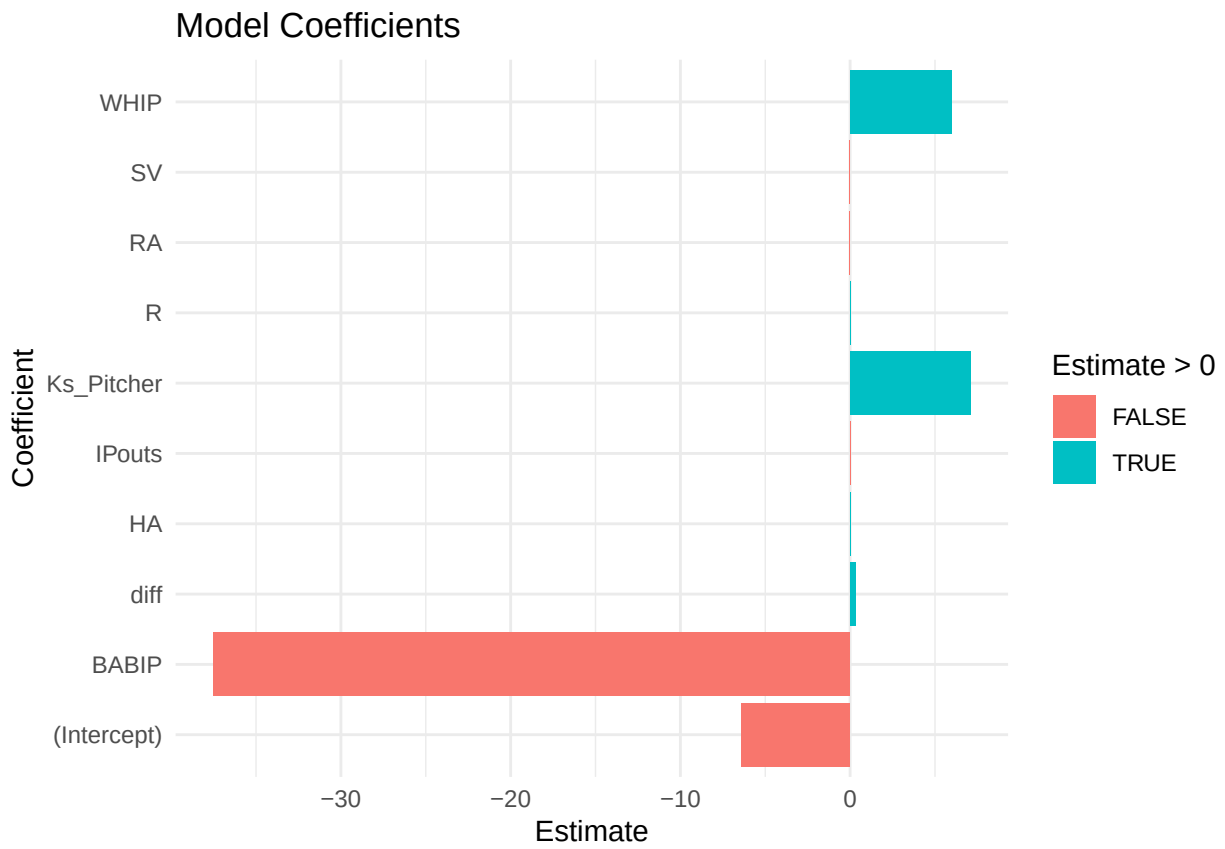
Variables R, RA, and diff have very small p-values, suggesting these predictors are statistically significant in relation to DivWin.

Null deviance: 434.75 on 439 degrees of freedom Residual deviance: 210.17 on 430 degrees of freedom

This indicates that the model with predictor variables provides a better fit to the data compared to the null model (intercept-only model), as shown by the reduction in deviance from 434.75 to 210.17.



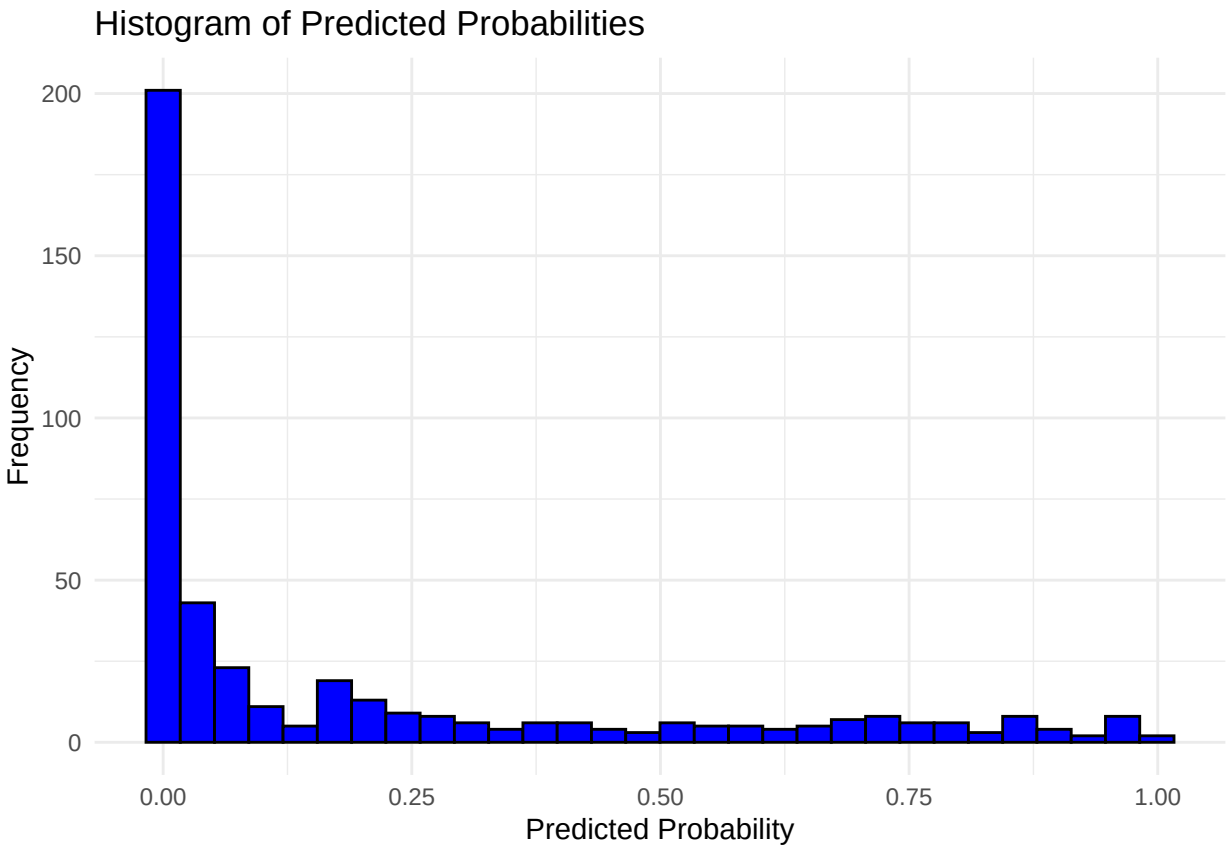
Residuals randomly distributed around y=0.



Each bar in the plot represents the estimated effect of one predictor variable on the log odds of the outcome (DivWin), holding all other variables constant.

Positive Coefficients: Variables with positive estimates (WHIP, diff) increase the log odds of the outcome. In your plot, these are shown with bars extending to the right of the zero line.

Negative Coefficients: Variables with negative estimates (BABIP, SV, RA) decrease the log odds of the outcome. These are shown with bars extending to the left.



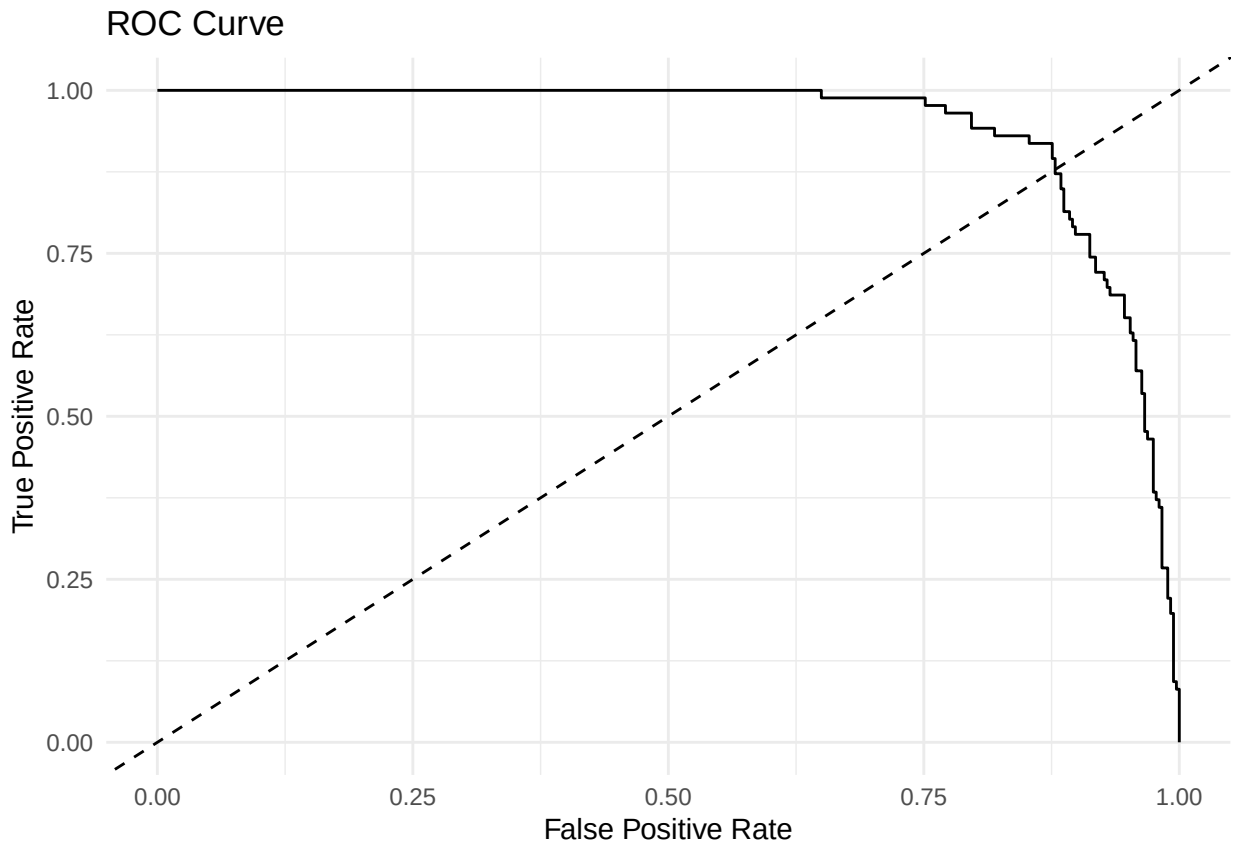
The shape of the histogram shows the distribution of predicted probabilities for DivWin. There is unimodal distribution heavily skewed towards one side. This might indicate that my model is uncertain about its predictions or biased towards one class.

The spread of the histogram (ranging from 0 to 1) shows how probabilities are distributed between the certain (close to 0 or 1) and uncertain (around 0.5) predictions.

Wider spreads generally indicate that the model is utilizing the available features effectively to differentiate between the outcomes. Conversely, a narrow spread centered around a particular value (like 0.5) might suggest that the model is not effectively distinguishing between the classes.

```
## Setting levels: control = N, case = Y
```

```
## Setting direction: controls < cases
```



The ROC (Receiver Operating Characteristic) curve is good for evaluating the performance of my logistic regression model, particularly in terms of its ability to distinguish between the two classes (e.g., "Y" and "N" in DivWin).

A high y-intercept in the ROC curve means that at the lowest threshold (just above the minimum predicted probability), the model is able to correctly identify a large proportion of the actual positives. This indicates strong sensitivity at lower thresholds.

The exponential decrease suggests that as I increase the threshold for predicting a positive class, the TPR decreases rapidly compared to the increase in FPR. This pattern often indicates that while the model is initially effective at identifying true positives, its ability to continue distinguishing positive cases becomes less effective as the threshold increases.

This type of ROC curve can sometimes indicate a model that performs well in identifying positive cases up to a certain point, after which its discriminative ability declines sharply. It might be particularly sensitive to changes in the threshold setting.