Major League Baseball Team Performance Analysis

Zack De Den, Mehedi Toufiqe, Isaac Clark

12-10-2021

1 Introduction

Our goal at K+1 is to determine the characteristics that affects a Major League Baseball team's winning percentage using the Lahman Baseball Database. The database, created and updated yearly by Sean Lahman, contains all Major League Baseball statistics from 1871 to present day. For the purpose of this study, we are constricting our analysis to the 'modern' era of baseball, years 1970 to 2019. *Data from years 1972, 1981, and 1994 were not included because the regular season was interrupted by player union strikes, thus are incomplete.*

Response Variable: Winning Percentage

Predictor Variables: Batting Average, Extra Base-hits per Game, Strikeouts per Game, Walks per Game, Hits Allowed per Game, Strikeouts Allowed per Game, Walks Allowed per Game, Earned Run Average

These predictors were chosen after careful thought to avoid extreme multicollinearity, as the most popular baseball statistics are computed using contextually similar information.

More succinctly, our goal is to determine which of these variables are relevant to predicting a team's winning percentage by analyzing the data from 1970 to 2019. We will also perform the same procedure on the corresponding decades within that time frame (1970s, 1980s, 1990, 2000s, amd 2010s).

2 Data Description

Description of Predictor Variables:

- *Batting Average:* $\frac{Hits}{At-Bats}$
  - represented by *BA*
- *Extra Base-hits per Game:* $\frac{Doubles+Triples+HomeRuns}{Games}$
  - represented by *XBPG*
- *Strikeouts per Game:* $\frac{Strikeouts}{Games}$
  - represented by *KPG* (strikeouts are colloquially known and recorded as the alphabetical symbol *K*)
- *Walks Per Game:* $\frac{Walks}{Games}$
  - represented by *BBPG* (walks are colloquially known and recorded as the alphabetical symbol *BB*)
- *Hits Allowed Per Game:* $\frac{HitsAllowed}{Games}$
  - represented by *HAPG*
- *Strikeouts Allowed Per Game:* $\frac{StrikeoutsAllowed}{Games}$

  - represented by $SOAPG$

- *Walks Allowed Per Game:* $\frac{WalksAllowed}{Games}$
  - represented by $BBAPG$

- *Earned Run Average:* $\frac{EarnedRuns \cdot 9}{InningsPitched}$
  - represented by ERA

Description of Response Variable:

- *Winning Percentage:* $\frac{Wins}{Games}$
  - represented by $WinPct$

3 Data Analysis

3.1 Cleaning

In order to ensure complete data and keep relevance in mind, the data analyzed was taken from years 1970 to 2019, minus the three years for reasons stated previously. The raw data was then selected and operated on to give us the rates of our predictor variables. No other cleaning steps were needed to be taken.

3.2 Model Construction and Validity

Our full fit regression model constructed as:

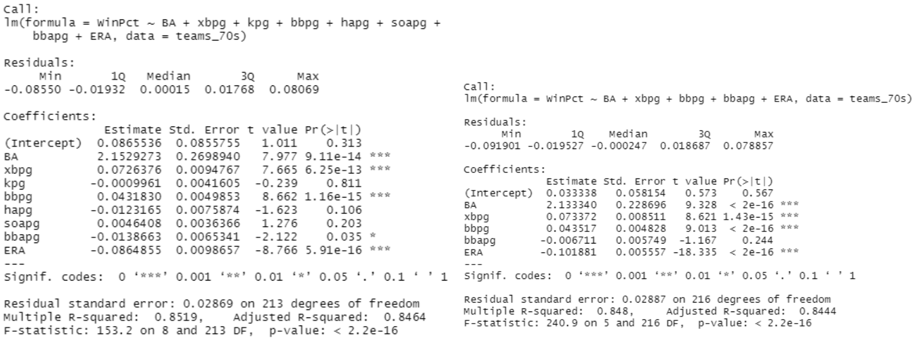
$$WinPct = BA + XBPG + KPG + BBPG + HAPG + SOAPG + BBAPG + ERA$$

After running a regression summary for our model, predictors with a p-value greater than 0.05 were removed from the model. The reduced model was then compared to the full model via an ANOVA check. Potentially influential, high leverage, and outlier points are identified and then removed. The same regression procedure is executed again on data with points removed.

4 The 1970s

The full fit regression model for the 1970s yielded a p-value greater than 0.05 for *KPG*, *HAPG*, and *SOAPG*, thus they were removed from the model. The reduced model yielded a slightly lower R-squared value than the full model (full: 0.8464, reduced: 0.8444). The reduced model passes ANOVA check with full model with a p-value of 0.1309. No other transformations on data was necessary, as residuals are randomly distributed between -3 and 3 in a horizontal band.

There were 11 highly influential points, 7 high leverage points, and one outlier point identified. After removing the influential points, leverage points, and outlier, the full fit regression model yielded a p-value greater than 0.05 for *KPG*, *HAPG*, *SOAPG*, and *BBAPG*, thus they were removed and the reduced model was executed again. The reduced model yielded a larger R-squared value of 0.8623. The ANOVA check for models with points removed yielded a p-value of 0.4071.



```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + soapg +
    bbapg + ERA, data = teams_70s)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.08550 -0.01932  0.00015  0.01768  0.08069 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.0865536  0.0855755   1.011    0.313    
BA           2.1529273  0.2698940   7.977 9.11e-14 ***
xbpg         0.0726376  0.0094767   7.665 6.25e-13 ***
kpg         -0.0009961  0.0041605  -0.239    0.811    
bbpg         0.0431830  0.0049853   8.662 1.16e-15 ***
hapg        -0.0123165  0.0075874  -1.623    0.106    
soapg        0.0046408  0.0036366   1.276    0.203    
bbapg       -0.0138663  0.0065341  -2.122    0.035 *  
ERA         -0.0864855  0.0098657  -8.766 5.91e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02869 on 213 degrees of freedom
Multiple R-squared:  0.8519,	Adjusted R-squared:  0.8464 
F-statistic: 153.2 on 8 and 213 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = WinPct ~ BA + xbpg + bbpg + bbapg + ERA, data = teams_70s)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.091901 -0.019527 -0.000247  0.018687  0.078857 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.033338   0.058154   0.573    0.567    
BA           2.133340   0.228696   9.328  < 2e-16 ***
xbpg         0.073372   0.008511   8.621 1.43e-15 ***
bbpg         0.043517   0.004828   9.013  < 2e-16 ***
bbapg       -0.006711   0.005749  -1.167    0.244    
ERA         -0.101881   0.005557 -18.335  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02887 on 216 degrees of freedom
Multiple R-squared:  0.848,	Adjusted R-squared:  0.8444 
F-statistic: 240.9 on 5 and 216 DF,  p-value: < 2.2e-16
```

Figure 1: Full and Reduced Models

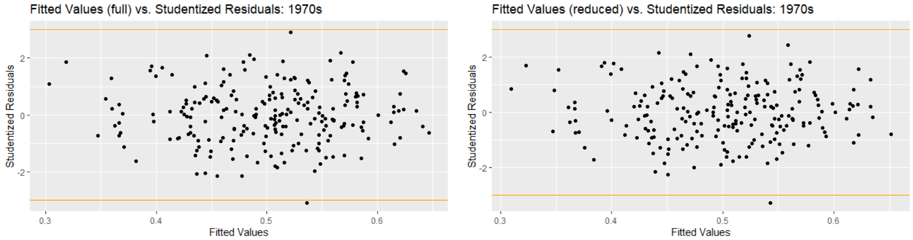


Figure 2: Full and Reduced Residual Plots

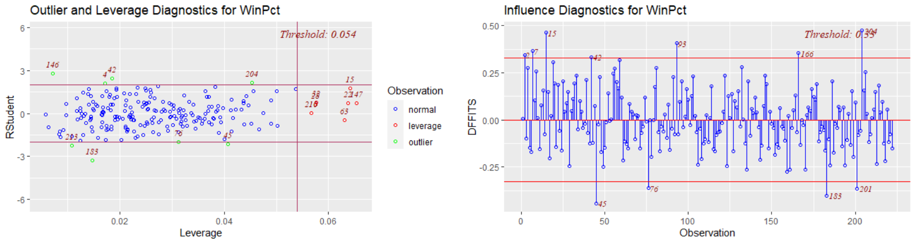
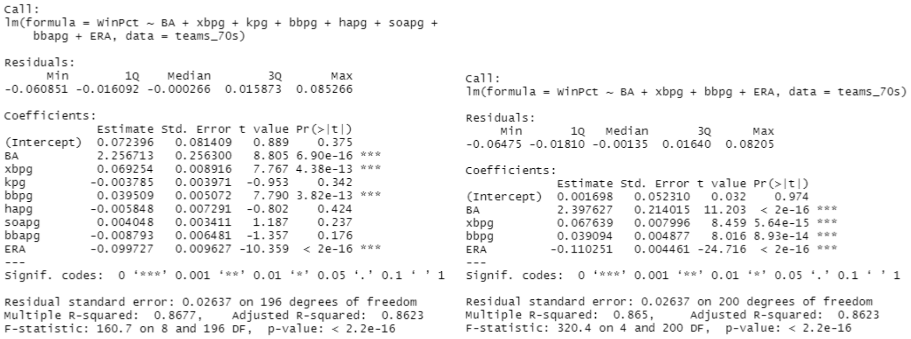


Figure 3: Leverage and Influence Plots



```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + soapg + 
    bbapg + ERA, data = teams_70s)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.060851 -0.016092 -0.000266  0.015873  0.085266 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.072396   0.081409   0.889    0.375    
BA           2.256713   0.256300   8.805 6.90e-16 ***
xbpg         0.069254   0.008916   7.767 4.38e-13 ***
kpg         -0.003785   0.003971  -0.953    0.342    
bbpg         0.039509   0.005072   7.790 3.82e-13 ***
hapg        -0.005848   0.007291  -0.802    0.424    
soapg        0.004048   0.003411   1.187    0.237    
bbapg       -0.008793   0.006481  -1.357    0.176    
ERA         -0.099727   0.009627 -10.359  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02637 on 196 degrees of freedom
Multiple R-squared:  0.8677,	Adjusted R-squared:  0.8623 
F-statistic: 160.7 on 8 and 196 DF,  p-value: < 2.2e-16
```

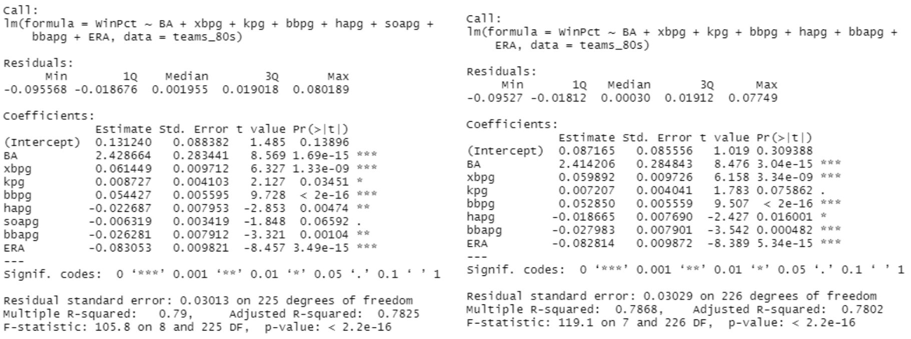
```
Call:
lm(formula = WinPct ~ BA + xbpg + bbpg + ERA, data = teams_70s)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.06475 -0.01810 -0.00135  0.01640  0.08205 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.001698   0.052310   0.032    0.974    
BA           2.397627   0.214015  11.203  < 2e-16 ***
xbpg         0.067639   0.007996   8.459 5.64e-15 ***
bbpg         0.039094   0.004877   8.016 8.93e-14 ***
ERA         -0.110251   0.004461 -24.716  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02637 on 200 degrees of freedom
Multiple R-squared:  0.865,	Adjusted R-squared:  0.8623 
F-statistic: 320.4 on 4 and 200 DF,  p-value: < 2.2e-16
```

Figure 4: Full and Reduced Models after point removal

5 The 1980s

The full fit regression model for the 1980s yielded a p-value greater than 0.05 for *SOAPG*, thus it was removed from the model. The p-value of *SOAPG* is 0.06592, which is pretty close to 0.05; this suggests some influential measures in regards to this predictor. The reduced model yielded a slightly lower R-squared value than the full model (full: 0.7825, reduced: 0.7802). The reduced model barely passes ANOVA check with full model with a p-value of 0.06592. No other transformations on data was necessary, as residuals are randomly distributed between -3 and 3 in a horizontal band.

There were 18 highly influential points, 7 high leverage points, and one outlier point identified. After removing the influential points, leverage points, and outlier, the full fit regression model yielded a p-value greater than 0.05 for *SOAPG*, thus it was removed and the reduced model was executed again. The p-value of *SOAPG* is 0.06141, which is pretty close to 0.05 once again. The reduced model yielded a larger R-squared value of 0.8454.



```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + soapg +
    bbapg + ERA, data = teams_80s)

Residuals:
      Min        1Q    Median        3Q       Max
-0.095568 -0.018676  0.001955  0.019018  0.080189

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.131240   0.088382   1.485  0.13896
BA           2.428664   0.283441   8.569 1.69e-15 ***
xbpg         0.061449   0.009712   6.327 1.33e-09 ***
kpg          0.008727   0.004103   2.127  0.03451 *
bbpg         0.054427   0.005595   9.728  < 2e-16 ***
hapg        -0.022687   0.007953  -2.853  0.00474 **
soapg       -0.006319   0.003419  -1.848  0.06592 .
bbapg       -0.026281   0.007912  -3.321  0.00104 **
ERA         -0.083053   0.009821  -8.457 3.49e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03013 on 225 degrees of freedom
Multiple R-squared:   0.79,     Adjusted R-squared:  0.7825
F-statistic: 105.8 on 8 and 225 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + bbapg +
    ERA, data = teams_80s)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09527 -0.01812  0.00030  0.01912  0.07749

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.087165   0.085556   1.019 0.309388
BA           2.414206   0.284843   8.476 3.04e-15 ***
xbpg         0.059892   0.009726   6.158 3.34e-09 ***
kpg          0.007207   0.004041   1.783 0.075862 .
bbpg         0.052850   0.005559   9.507  < 2e-16 ***
hapg        -0.018665   0.007690  -2.427 0.016001 *
bbapg       -0.027983   0.007901  -3.542 0.000482 ***
ERA         -0.082814   0.009872  -8.389 5.34e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03029 on 226 degrees of freedom
Multiple R-squared:  0.7868,    Adjusted R-squared:  0.7802
F-statistic: 119.1 on 7 and 226 DF,  p-value: < 2.2e-16
```

Figure 5: Full and Reduced Models

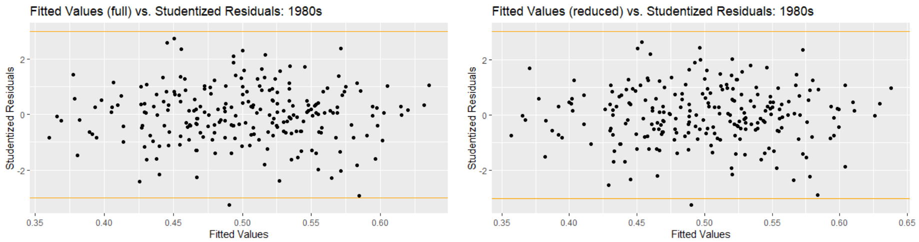


Figure 6: Full and Reduced Residual Plots

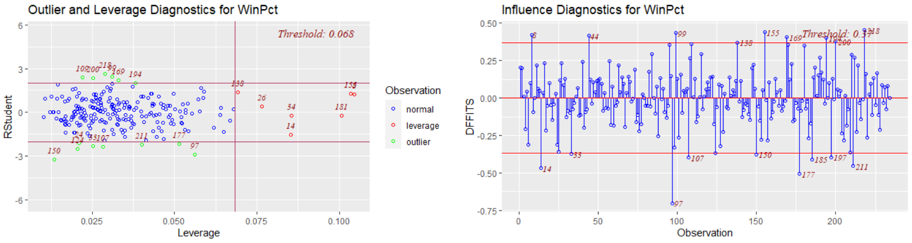
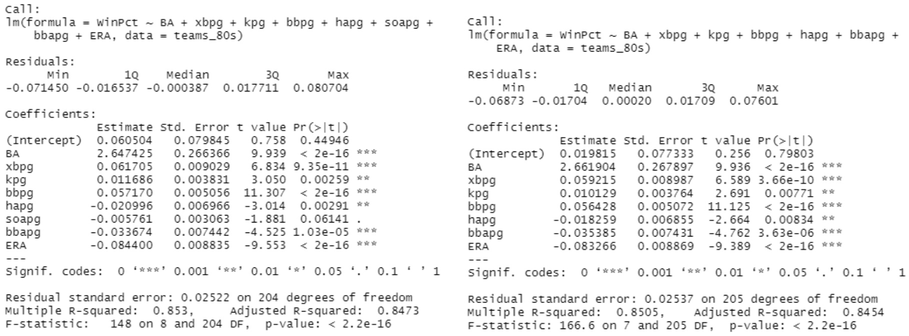


Figure 7: Leverage and Influence Plots



```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + soapg +
    bbapg + ERA, data = teams_80s)

Residuals:
      Min        1Q    Median        3Q       Max
-0.071450 -0.016537 -0.000387  0.017711  0.080704

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.060504   0.079845   0.758  0.44946
BA           2.647425   0.266366   9.939  < 2e-16 ***
xbpg         0.061705   0.009029   6.834 9.35e-11 ***
kpg          0.011686   0.003831   3.050  0.00259 **
bbpg         0.057170   0.005056  11.307  < 2e-16 ***
hapg        -0.020996   0.006966  -3.014  0.00291 **
soapg       -0.005761   0.003063  -1.881  0.06141 .
bbapg       -0.033674   0.007442  -4.525 1.03e-05 ***
ERA         -0.084400   0.008835  -9.553  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02522 on 204 degrees of freedom
Multiple R-squared:  0.853,     Adjusted R-squared:  0.8473
F-statistic:   148 on 8 and 204 DF,  p-value: < 2.2e-16
```

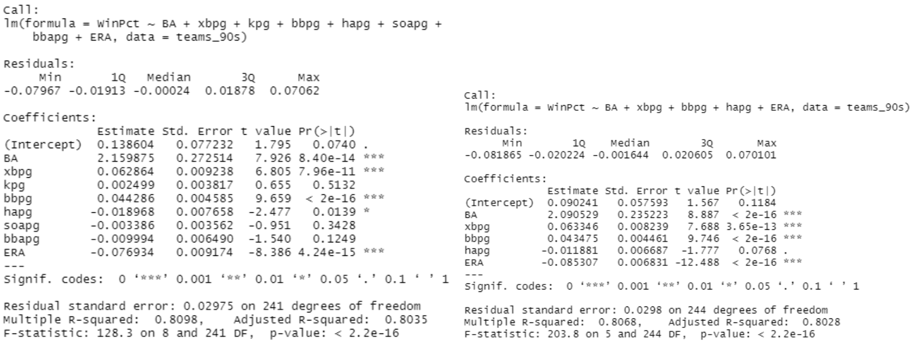
```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + bbapg +
    ERA, data = teams_80s)

Residuals:
     Min       1Q   Median       3Q      Max
-0.06873 -0.01704  0.00020  0.01709  0.07601

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.019815   0.077333   0.256  0.79803
BA           2.661904   0.267897   9.936  < 2e-16 ***
xbpg         0.059215   0.008987   6.589 3.66e-10 ***
kpg          0.010129   0.003764   2.691  0.00771 **
bbpg         0.056428   0.005072  11.125  < 2e-16 ***
hapg        -0.018259   0.006855  -2.664  0.00834 **
bbapg       -0.035385   0.007431  -4.762 3.63e-06 ***
ERA         -0.083266   0.008869  -9.389  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02537 on 205 degrees of freedom
Multiple R-squared:  0.8505,    Adjusted R-squared:  0.8454
F-statistic: 166.6 on 7 and 205 DF,  p-value: < 2.2e-16
```

Figure 8: Full and Reduced Models after point removal

6 The 1990s

The full fit regression model for the 1990s yielded a p-value greater than 0.05 for *KPG*, *SOAPG*, and *BBAPG*, thus they were removed from the model. The reduced model yielded a slightly lower R-squared value than the full model (full: 0.8035, reduced: 0.8028). The reduced model passes ANOVA check with full model with a p-value of 0.28. No other transformations on data was necessary, as residuals are randomly distributed between -3 and 3 in a horizontal band.

There were 18 highly influential points and 12 high leverage points identified. After removing the influential points and leverage points, the full fit regression model yielded a p-value greater than 0.05 for *KPG*, *HAPG*, *SOAPG*, and *BBAPG*, thus they were removed and the reduced model was executed again. The p-values of *HAPG* and *BBAPG* are 0.0510 and 0.0576 respectively, which are pretty close to 0.05. The reduced model yielded a larger R-squared value of 0.821. The ANOVA check for models with points removed yielded a p-value of 0.259.



```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + soapg +
    bbapg + ERA, data = teams_90s)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.07967 -0.01913 -0.00024  0.01878  0.07062 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.138604   0.077232   1.795   0.0740 .  
BA           2.159875   0.272514   7.926 8.40e-14 ***
xbpg         0.062864   0.009238   6.805 7.96e-11 ***
kpg          0.002499   0.003817   0.655   0.5132    
bbpg         0.044286   0.004585   9.659  < 2e-16 ***
hapg        -0.018968   0.007658  -2.477   0.0139 *  
soapg       -0.003386   0.003562  -0.951   0.3428    
bbapg       -0.009994   0.006490  -1.540   0.1249    
ERA         -0.076934   0.009174  -8.386 4.24e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02975 on 241 degrees of freedom
Multiple R-squared:  0.8098,    Adjusted R-squared:  0.8035 
F-statistic: 128.3 on 8 and 241 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = WinPct ~ BA + xbpg + bbpg + hapg + ERA, data = teams_90s)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.081865 -0.020224 -0.001644  0.020605  0.070101 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.090241   0.057593   1.567   0.1184    
BA           2.090529   0.235223   8.887  < 2e-16 ***
xbpg         0.063346   0.008239   7.688 3.65e-13 ***
bbpg         0.043475   0.004461   9.746  < 2e-16 ***
hapg        -0.011881   0.006687  -1.777   0.0768 .  
ERA         -0.085307   0.006831 -12.488  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0298 on 244 degrees of freedom
Multiple R-squared:  0.8068,    Adjusted R-squared:  0.8028 
F-statistic: 203.8 on 5 and 244 DF,  p-value: < 2.2e-16
```

Figure 9: Full and Reduced Models

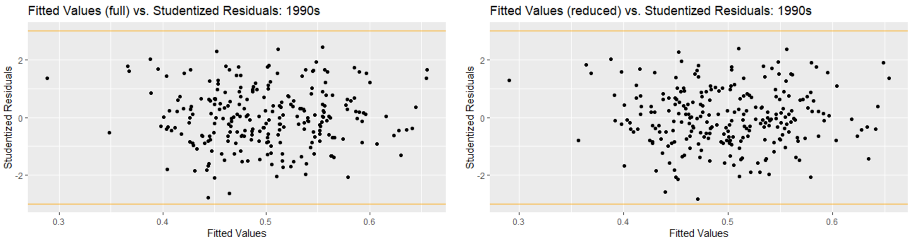


Figure 10: Full and Reduced Residual Plots

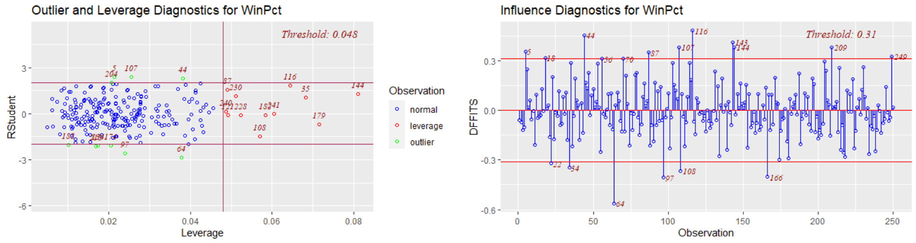
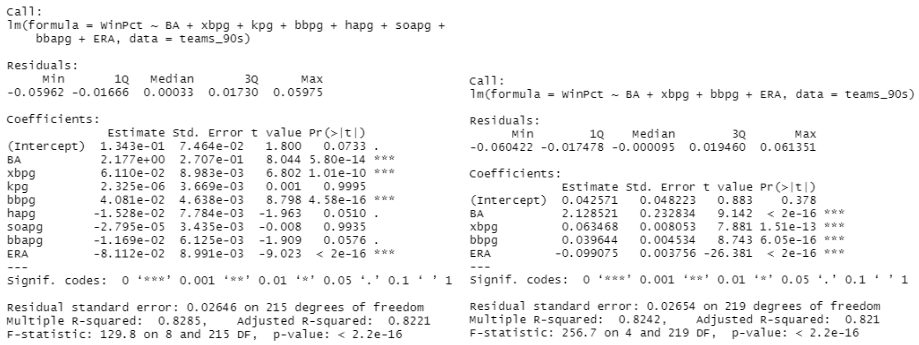


Figure 11: Leverage and Influence Plots



```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + soapg +
    bbapg + ERA, data = teams_90s)

Residuals:
     Min       1Q   Median       3Q      Max
-0.05962 -0.01666  0.00033  0.01730  0.05975

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.343e-01  7.464e-02   1.800   0.0733 .
BA           2.177e+00  2.707e-01   8.044 5.80e-14 ***
xbpg         6.110e-02  8.983e-03   6.802 1.01e-10 ***
kpg          2.325e-06  3.669e-03   0.001   0.9995
bbpg         4.081e-02  4.638e-03   8.798 4.58e-16 ***
hapg        -1.528e-02  7.784e-03  -1.963   0.0510 .
soapg       -2.795e-05  3.435e-03  -0.008   0.9935
bbapg       -1.169e-02  6.125e-03  -1.909   0.0576 .
ERA         -8.112e-02  8.991e-03  -9.023  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02646 on 215 degrees of freedom
Multiple R-squared:  0.8285,	Adjusted R-squared:  0.8221
F-statistic: 129.8 on 8 and 215 DF,  p-value: < 2.2e-16
```

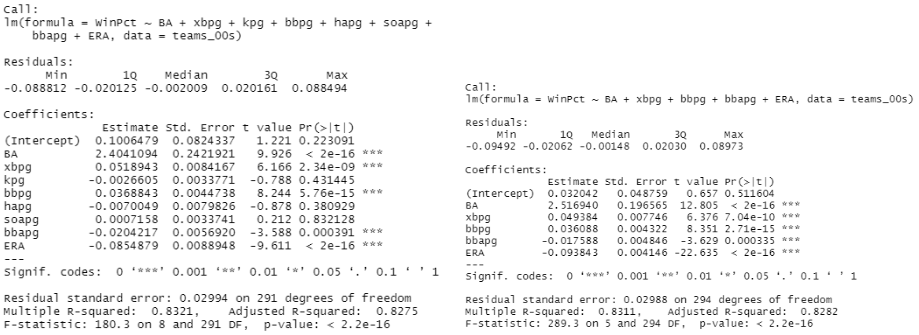
```
Call:
lm(formula = WinPct ~ BA + xbpg + bbpg + ERA, data = teams_90s)

Residuals:
      Min        1Q    Median        3Q       Max
-0.060422 -0.017478 -0.000095  0.019460  0.061351

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.042571   0.048223   0.883    0.378
BA           2.128521   0.232834   9.142  < 2e-16 ***
xbpg         0.063468   0.008053   7.881 1.51e-13 ***
bbpg         0.039644   0.004534   8.743 6.05e-16 ***
ERA         -0.099075   0.003756 -26.381  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02654 on 219 degrees of freedom
Multiple R-squared:  0.8242,	Adjusted R-squared:  0.821
F-statistic: 256.7 on 4 and 219 DF,  p-value: < 2.2e-16
```

Figure 12: Full and Reduced Models after point removal

7 The 2000s

The full fit regression model for the 2000s yielded a p-value greater than 0.05 for *KPG*, *HAPG*, and *SOAPG*, thus they were removed from the model. The reduced model yielded a slightly higher R-squared value than the full model (full: 0.8275, reduced: 0.8282). The reduced model passes ANOVA check with full model with p-value of 0.6211. No other transformations on data was necessary, as residuals are randomly distributed between -3 and 3 in a horizontal band.

There were 18 highly influential points, 15 high leverage points, and 2 outlier points identified. After removing the influential points, leverage points, and outliers, the full fit regression model yielded a p-value greater than 0.05 for *KPG*, *HAPG*, and *SOAPG*, thus they were removed and the reduced model was executed again. The reduced model yielded a larger R-squared value of 0.8568. The ANOVA check for models with points removed yielded a p-value of 0.4352.



```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + soapg + 
    bbapg + ERA, data = teams_00s)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.088812 -0.020125 -0.002009  0.020161  0.088494 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.1006479  0.0824337   1.221 0.223091    
BA           2.4041094  0.2421921   9.926  < 2e-16 ***
xbpg         0.0518943  0.0084167   6.166 2.34e-09 ***
kpg         -0.0026605  0.0033771  -0.788 0.431445    
bbpg         0.0368843  0.0044738   8.244 5.76e-15 ***
hapg        -0.0070049  0.0079826  -0.878 0.380929    
soapg        0.0007158  0.0033741   0.212 0.832128    
bbapg       -0.0204217  0.0056920  -3.588 0.000391 ***
ERA         -0.0854879  0.0088948  -9.611  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02994 on 291 degrees of freedom
Multiple R-squared:  0.8321,	Adjusted R-squared:  0.8275 
F-statistic: 180.3 on 8 and 291 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = WinPct ~ BA + xbpg + bbpg + bbapg + ERA, data = teams_00s)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.09492 -0.02062 -0.00148  0.02030  0.08973 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.032042   0.048759   0.657 0.511604    
BA           2.516940   0.196565  12.805  < 2e-16 ***
xbpg         0.049384   0.007746   6.376 7.04e-10 ***
bbpg         0.036088   0.004322   8.351 2.71e-15 ***
bbapg       -0.017588   0.004846  -3.629 0.000335 ***
ERA         -0.093843   0.004146 -22.635  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02988 on 294 degrees of freedom
Multiple R-squared:  0.8311,	Adjusted R-squared:  0.8282 
F-statistic: 289.3 on 5 and 294 DF,  p-value: < 2.2e-16
```

Figure 13: Full and Reduced Models

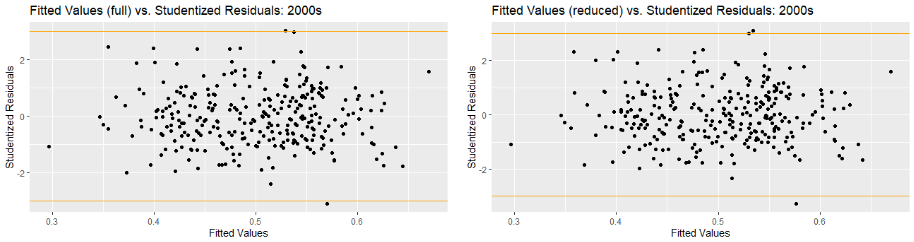


Figure 14: Full and Reduced Residual Plots

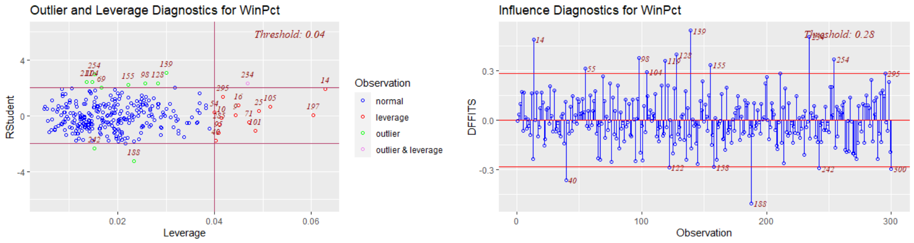
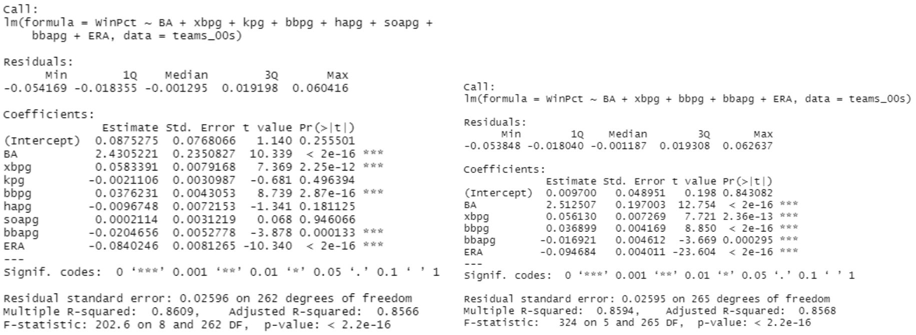


Figure 15: Leverage and Influence Plots



```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + soapg +
    bbapg + ERA, data = teams_00s)

Residuals:
      Min        1Q    Median        3Q       Max
-0.054169 -0.018355 -0.001295  0.019198  0.060416

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0875275  0.0768066   1.140 0.255501
BA           2.4305221  0.2350827  10.339  < 2e-16 ***
xbpg         0.0583391  0.0079168   7.369 2.25e-12 ***
kpg         -0.0021106  0.0030987  -0.681 0.496394
bbpg         0.0376231  0.0043053   8.739 2.87e-16 ***
hapg        -0.0096748  0.0072153  -1.341 0.181125
soapg        0.0002114  0.0031219   0.068 0.946066
bbapg       -0.0204656  0.0052778  -3.878 0.000133 ***
ERA         -0.0840246  0.0081265 -10.340  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02596 on 262 degrees of freedom
Multiple R-squared:  0.8609,    Adjusted R-squared:  0.8566
F-statistic: 202.6 on 8 and 262 DF,  p-value: < 2.2e-16
```

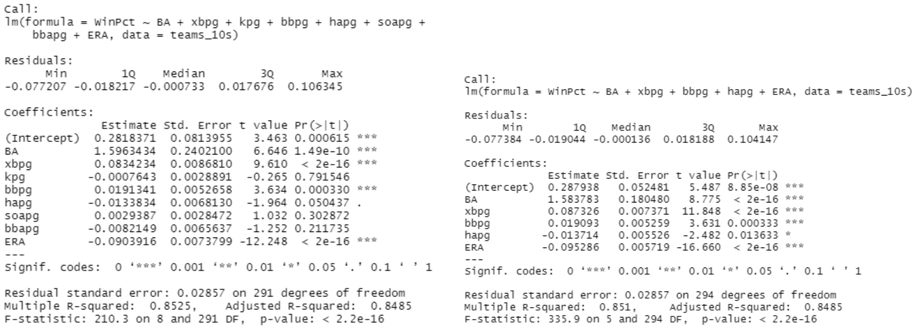
```
Call:
lm(formula = WinPct ~ BA + xbpg + bbpg + bbapg + ERA, data = teams_00s)

Residuals:
      Min        1Q    Median        3Q       Max
-0.053848 -0.018040 -0.001187  0.019308  0.062637

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.009700   0.048951   0.198 0.843082
BA           2.512507   0.197003  12.754  < 2e-16 ***
xbpg         0.056130   0.007269   7.721 2.36e-13 ***
bbpg         0.036899   0.004169   8.850  < 2e-16 ***
bbapg       -0.016921   0.004612  -3.669 0.000295 ***
ERA         -0.094684   0.004011 -23.604  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02595 on 265 degrees of freedom
Multiple R-squared:  0.8594,    Adjusted R-squared:  0.8568
F-statistic:   324 on 5 and 265 DF,  p-value: < 2.2e-16
```

Figure 16: Full and Reduced Models after point removal

8 The 2010s

The full fit regression model for the 2010s yielded a p-value greater than 0.05 for *KPG*, *SOAPG*, and *BBAPG*, thus they were removed from the model. The predictor *HAPG* returned a p-value of 0.050437; this is extremely close to 0.05, so we decided to keep it in our model. The full and the reduced model yielded an R-squared value of 0.8485. The reduced model passes ANOVA check with full model with p-value of 0.4049. No other transformations on data was necessary, as residuals are randomly distributed between -3 and 3 in a horizontal band.

There were 23 highly influential points, 29 high leverage points, and one outlier point identified. After removing the influential points, leverage points, and outlier, the full fit regression model yielded a p-value greater than 0.05 for *KPG*, *SOAPG*, and *BBAPG*, thus they were removed and the reduced model was executed again. The reduced model yielded a larger R-squared value of 0.867. The ANOVA check for models with points removed yielded a p-value of 0.3177.



```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + soapg +
    bbapg + ERA, data = teams_10s)

Residuals:
      Min        1Q    Median        3Q       Max
-0.077207 -0.018217 -0.000733  0.017676  0.106345

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2818371  0.0813955   3.463 0.000615 ***
BA           1.5963434  0.2402100   6.646 1.49e-10 ***
xbpg         0.0834234  0.0086810   9.610  < 2e-16 ***
kpg         -0.0007643  0.0028891  -0.265 0.791546
bbpg         0.0191341  0.0052658   3.634 0.000330 ***
hapg        -0.0133834  0.0068130  -1.964 0.050437 .
soapg        0.0029387  0.0028472   1.032 0.302872
bbapg       -0.0082149  0.0065637  -1.252 0.211735
ERA         -0.0903916  0.0073799 -12.248  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02857 on 291 degrees of freedom
Multiple R-squared:  0.8525,    Adjusted R-squared:  0.8485
F-statistic: 210.3 on 8 and 291 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = WinPct ~ BA + xbpg + bbpg + hapg + ERA, data = teams_10s)

Residuals:
      Min        1Q    Median        3Q       Max
-0.077384 -0.019044 -0.000136  0.018188  0.104147

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.287938   0.052481   5.487 8.85e-08 ***
BA           1.583783   0.180480   8.775  < 2e-16 ***
xbpg         0.087326   0.007371  11.848  < 2e-16 ***
bbpg         0.019093   0.005259   3.631 0.000333 ***
hapg        -0.013714   0.005526  -2.482 0.013633 *
ERA         -0.095286   0.005719 -16.660  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02857 on 294 degrees of freedom
Multiple R-squared:  0.851,     Adjusted R-squared:  0.8485
F-statistic: 335.9 on 5 and 294 DF,  p-value: < 2.2e-16
```

Figure 17: Full and Reduced Models

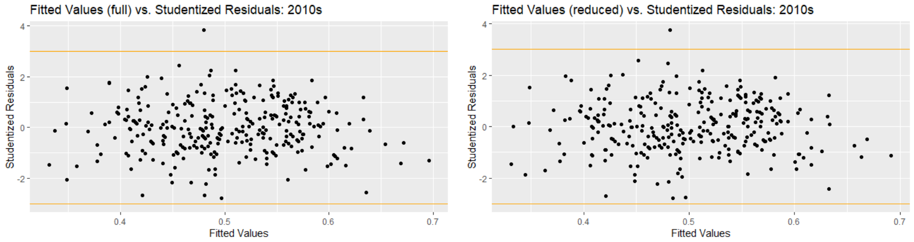


Figure 18: Full and Reduced Residual Plots

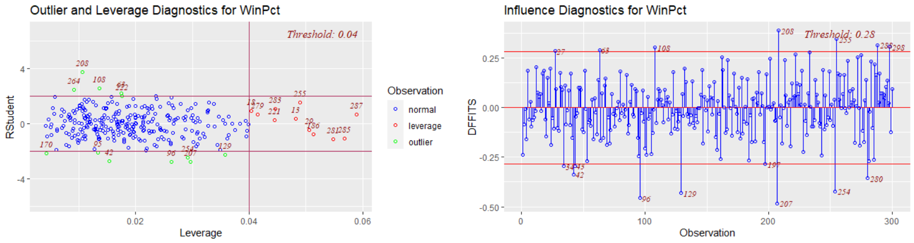
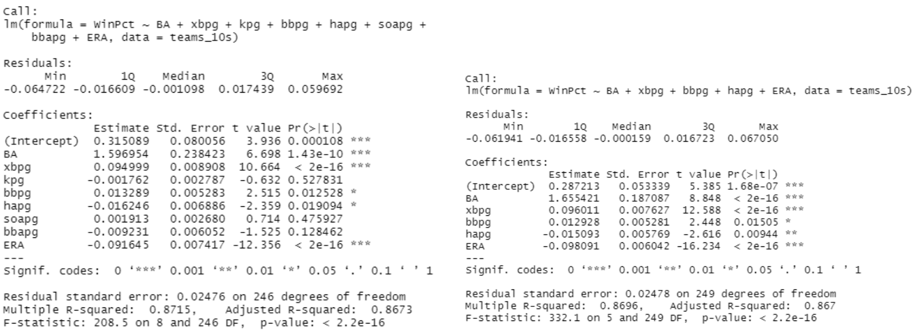


Figure 19: Leverage and Influence Plots



```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + soapg +
    bbapg + ERA, data = teams_10s)

Residuals:
      Min        1Q    Median        3Q       Max
-0.064722 -0.016609 -0.001098  0.017439  0.059692

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.315089   0.080056   3.936 0.000108 ***
BA           1.596954   0.238423   6.698 1.43e-10 ***
xbpg         0.094999   0.008908  10.664  < 2e-16 ***
kpg         -0.001762   0.002787  -0.632 0.527831
bbpg         0.013289   0.005283   2.515 0.012528 *
hapg        -0.016246   0.006886  -2.359 0.019094 *
soapg        0.001913   0.002680   0.714 0.475927
bbapg       -0.009231   0.006052  -1.525 0.128462
ERA         -0.091645   0.007417 -12.356  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02476 on 246 degrees of freedom
Multiple R-squared:  0.8715,	Adjusted R-squared:  0.8673
F-statistic: 208.5 on 8 and 246 DF,  p-value: < 2.2e-16
```

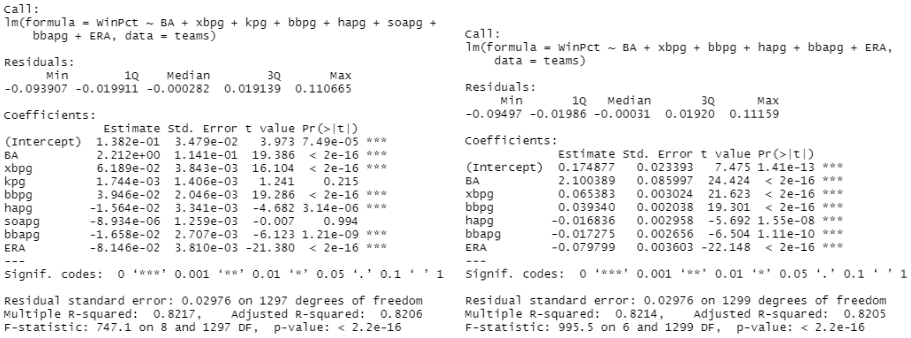
```
Call:
lm(formula = WinPct ~ BA + xbpg + bbpg + hapg + ERA, data = teams_10s)

Residuals:
      Min        1Q    Median        3Q       Max
-0.061941 -0.016558 -0.000159  0.016723  0.067050

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.287213   0.053339   5.385 1.68e-07 ***
BA           1.655421   0.187087   8.848  < 2e-16 ***
xbpg         0.096011   0.007627  12.588  < 2e-16 ***
bbpg         0.012928   0.005281   2.448  0.01505 *
hapg        -0.015093   0.005769  -2.616  0.00944 **
ERA         -0.098091   0.006042 -16.234  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02478 on 249 degrees of freedom
Multiple R-squared:  0.8696,	Adjusted R-squared:  0.867
F-statistic: 332.1 on 5 and 249 DF,  p-value: < 2.2e-16
```

Figure 20: Full and Reduced Models after point removal

9 From 1970 to 2019

The full fit regression model for 1970 to 2019 yielded a p-value greater than 0.05 for *KPG* and *SOAPG*, thus they were removed from the model. Both the full and reduced model yielded an R-squared value of 0.820. The reduced model passes ANOVA check with full model with p-value of 0.3189. No other transformations on data was necessary, as residuals are randomly distributed between -3 and 3 in a horizontal band.

There were 71 highly influential points, 44 high leverage points, and 3 outlier points identified. After removing the influential points, leverage points, and outliers, the full fit regression model yielded a p-value greater than 0.05 for *KPG* and *SOAPG*, thus they were removed and the reduced model was executed again. The reduced model yielded a larger R-squared value of 0.848. The ANOVA check for models with points removed yielded a p-value of 0.682.



```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + soapg + 
    bbapg + ERA, data = teams)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.093907 -0.019911 -0.000282  0.019139  0.110665 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.382e-01  3.479e-02   3.973 7.49e-05 ***
BA           2.212e+00  1.141e-01  19.386  < 2e-16 ***
xbpg         6.189e-02  3.843e-03  16.104  < 2e-16 ***
kpg          1.744e-03  1.406e-03   1.241    0.215    
bbpg         3.946e-02  2.046e-03  19.286  < 2e-16 ***
hapg        -1.564e-02  3.341e-03  -4.682 3.14e-06 ***
soapg       -8.934e-06  1.259e-03  -0.007    0.994    
bbapg       -1.658e-02  2.707e-03  -6.123 1.21e-09 ***
ERA         -8.146e-02  3.810e-03 -21.380  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02976 on 1297 degrees of freedom
Multiple R-squared:  0.8217,	Adjusted R-squared:  0.8206 
F-statistic: 747.1 on 8 and 1297 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = WinPct ~ BA + xbpg + bbpg + hapg + bbapg + ERA, 
    data = teams)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.09497 -0.01986 -0.00031  0.01920  0.11159 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.174877   0.023393   7.475 1.41e-13 ***
BA           2.100389   0.085997  24.424  < 2e-16 ***
xbpg         0.065383   0.003024  21.623  < 2e-16 ***
bbpg         0.039340   0.002038  19.301  < 2e-16 ***
hapg        -0.016836   0.002958  -5.692 1.55e-08 ***
bbapg       -0.017275   0.002656  -6.504 1.11e-10 ***
ERA         -0.079799   0.003603 -22.148  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02976 on 1299 degrees of freedom
Multiple R-squared:  0.8214,	Adjusted R-squared:  0.8205 
F-statistic: 995.5 on 6 and 1299 DF,  p-value: < 2.2e-16
```

Figure 21: Full and Reduced Models

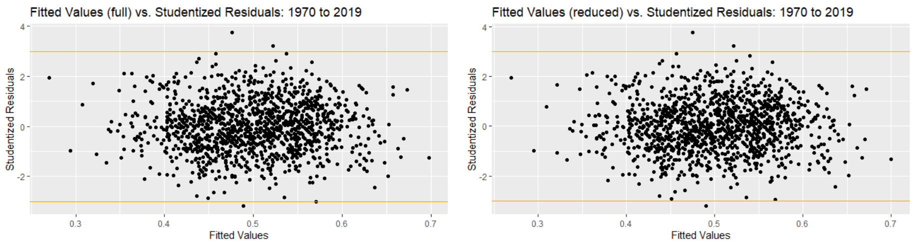


Figure 22: Full and Reduced Residual Plots

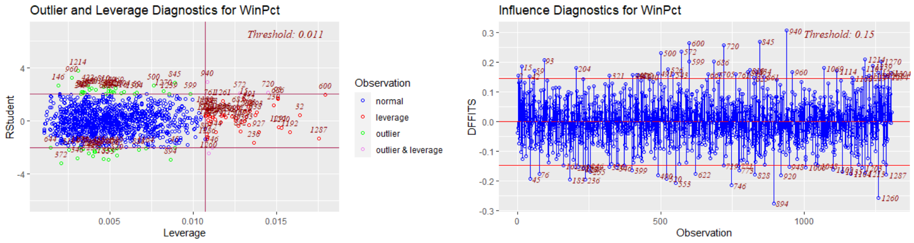
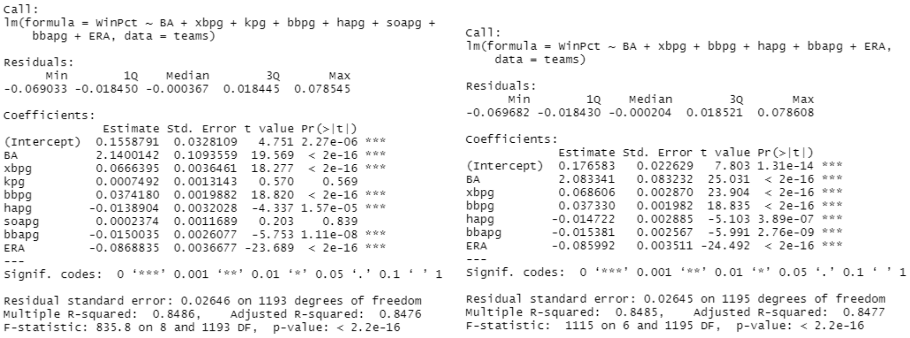


Figure 23: Leverage and Influence Plots



```
Call:
lm(formula = WinPct ~ BA + xbpg + kpg + bbpg + hapg + soapg + 
    bbapg + ERA, data = teams)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.069033 -0.018450 -0.000367  0.018445  0.078545 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.1558791  0.0328109   4.751 2.27e-06 ***
BA           2.1400142  0.1093559  19.569  < 2e-16 ***
xbpg         0.0666395  0.0036461  18.277  < 2e-16 ***
kpg          0.0007492  0.0013143   0.570    0.569    
bbpg         0.0374180  0.0019882  18.820  < 2e-16 ***
hapg        -0.0138904  0.0032028  -4.337 1.57e-05 ***
soapg        0.0002374  0.0011689   0.203    0.839    
bbapg       -0.0150035  0.0026077  -5.753 1.11e-08 ***
ERA         -0.0868835  0.0036677 -23.689  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02646 on 1193 degrees of freedom
Multiple R-squared:  0.8486,	Adjusted R-squared:  0.8476 
F-statistic: 835.8 on 8 and 1193 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = WinPct ~ BA + xbpg + bbpg + hapg + bbapg + ERA, 
    data = teams)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.069682 -0.018430 -0.000204  0.018521  0.078608 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.176583   0.022629   7.803 1.31e-14 ***
BA           2.083341   0.083232  25.031  < 2e-16 ***
xbpg         0.068606   0.002870  23.904  < 2e-16 ***
bbpg         0.037330   0.001982  18.835  < 2e-16 ***
hapg        -0.014722   0.002885  -5.103 3.89e-07 ***
bbapg       -0.015381   0.002567  -5.991 2.76e-09 ***
ERA         -0.085992   0.003511 -24.492  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02645 on 1195 degrees of freedom
Multiple R-squared:  0.8485,	Adjusted R-squared:  0.8477 
F-statistic:  1115 on 6 and 1195 DF,  p-value: < 2.2e-16
```

Figure 24: Full and Reduced Models after point removal

10 Conclusion

After looking at the reduced regression models from each decade and the entire time frame, we can see that the variables Batting Average, Extra Base-hits per Game, Walks per Game, and Earned Run Average appear on all reduced models. Overall, we can conclude that teams that can hit for high average and power, get on base via walks at a sustainable rate, and keep a low earned run average are more likely, to have a high winning percentage, thus putting them in a better position to make the playoffs.

The models from the 1980s and the 1990s are of particular note, as there were predictors very close to p-value of 0.05. For the 1980s, the predictor $SOAPG$ maybe be important in capturing a more complete picture of how teams performed in that decade. Anecdotally, the 1980s are known retroactively as a pitcher dominant decade, so analyzing more specific team pitching data may be necessary. For the 1990s, the predictors $HAPG$ and $BBAPG$ may also be important in the same way. The size of the league expanded from 26 to 30 during the 1990s, adding four new teams from 1993 to 1998. These expansion teams faced a steep entry barrier competition-wise, as they did not have the resources or talent that more established teams had built. This may account for the borderline nature of these predictors.

All predictors were chosen to keep extreme multicollinearity to a minimum. However, we acknowledge that given the nature of baseball statistics, multicollinearity may be unavoidable when choosing different predictors. Further comparisons and analysis can be done through other analytic means developed for baseball statistics, such as the Pythagorean Win-Loss Ratio or $wOBA$, weighted on-base average.