STAT 741 Final Project

Alex Clark

I. Research Topic

I will be attempting to predict the number of wins for a team in a single Major League Baseball season. Predicting wins is crucial for an MLB team to gauge where they think the team currently is and what targets they should be looking at acquiring. Two indicators that seems to stand out as predictors of wins will be runs scored (R) total and earned run average (ERA). The runs scored will attempt to predict the offenses contribution and conversely the ERA will attempt to predict the defenses contribution. My theory is the teams with a larger runs scored and lower ERA will have a larger number of wins, and teams with a lower runs scored and higher ERA will have a smaller number of wins.

II. Data Collection/Data Source

The data for my final project be using MLB data back to 1962. I am using back to 1962 because that is when the MLB went to a standard 162 game schedule for each team in both leagues. I will be using each team's Runs (R) total for the season and their Earned Run Average (ERA) for the season. These will be my 2 predictor variables, and wins will be my response variable. I have the data via Sean Lahman, who maintains a database of MLB stats back to the 1800s on his website. The data I am using will have 1,517 data points. The data does not need to be cleaned or transformed in anyway. I will be using R to conduct my analysis.

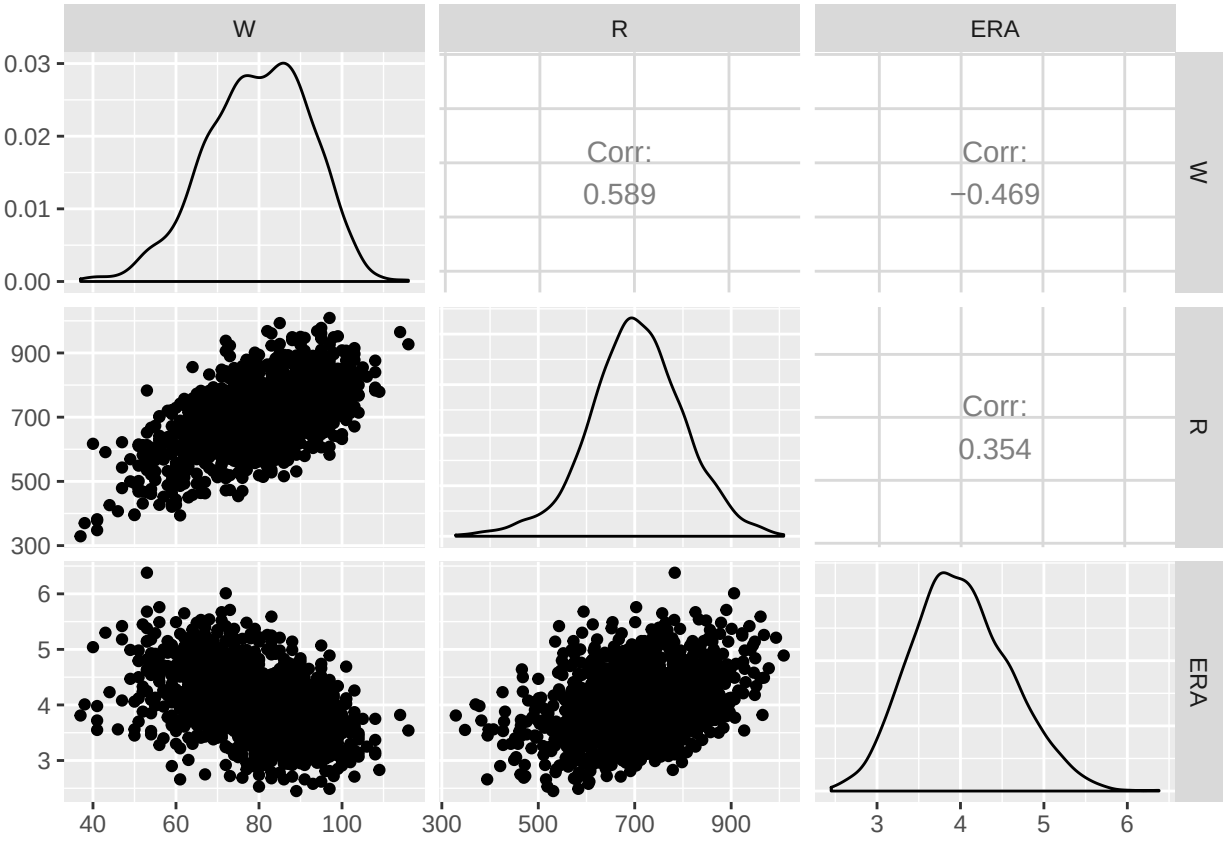
Data:

	W	R	ERA
1378	77	652	3.69
1379	76	707	4.22
1380	85	707	3.73

Summary of Data:

W	R	ERA
Min. : 37.00	Min. : 329.0	Min. :2.450
1st Qu.: 71.00	1st Qu.: 642.0	1st Qu.:3.600
Median : 80.00	Median : 704.0	Median :3.980
Mean : 79.79	Mean : 703.4	Mean :4.011
3rd Qu.: 89.00	3rd Qu.: 767.0	3rd Qu.:4.390
Max. :116.00	Max. :1009.0	Max. :6.380

Plot - Correlation of Data:



Overall, the data looks to have a linear relationship and is properly correlated to my expectations. Runs should be positively correlated to wins, and ERA should be negatively correlated to wins.

III. Method of Analysis

The model will be using R and ERA to predict W. The assumption is taken that every team will play at least 162 games, some teams may play 163 games depending on tie-breakers. The model will only be looking at MLB data, this does not include minor leagues, international leagues or international tournaments. The model will be a linear regression model. To check the validity of the model I will be looking at R2, residual plots, QQ plot, confidence intervals, and also checking the model for multicollinearity.

```
wins.lm <- lm(W ~ R + ERA, data = teams.subset)
wins.lm
```

```
##
## Call:
## lm(formula = W ~ R + ERA, data = teams.subset)
##
## Coefficients:
## (Intercept)            R          ERA
##      69.023        0.108      -16.252
```

Linear Regression Equation:

$\hat{y} = 69.02 + 0.11_R$ -16.25_{ERA}

Summary of Model:

```
summary(wins.lm)
```

```
##
## Call:
## lm(formula = W ~ R + ERA, data = teams.subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.7559  -2.9889   0.0566   2.9205  14.4039
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.023109   0.967736   71.32   <2e-16 ***
## R             0.107965   0.001228   87.91   <2e-16 ***
## ERA         -16.251787   0.206068  -78.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.412 on 1515 degrees of freedom
## Multiple R-squared:  0.8721, Adjusted R-squared:  0.872
## F-statistic:  5167 on 2 and 1515 DF,  p-value: < 2.2e-16
```

Runs t-value: 87.91, Runs p-value: 0

ERA t-value: -78.87, ERA p-value: 0

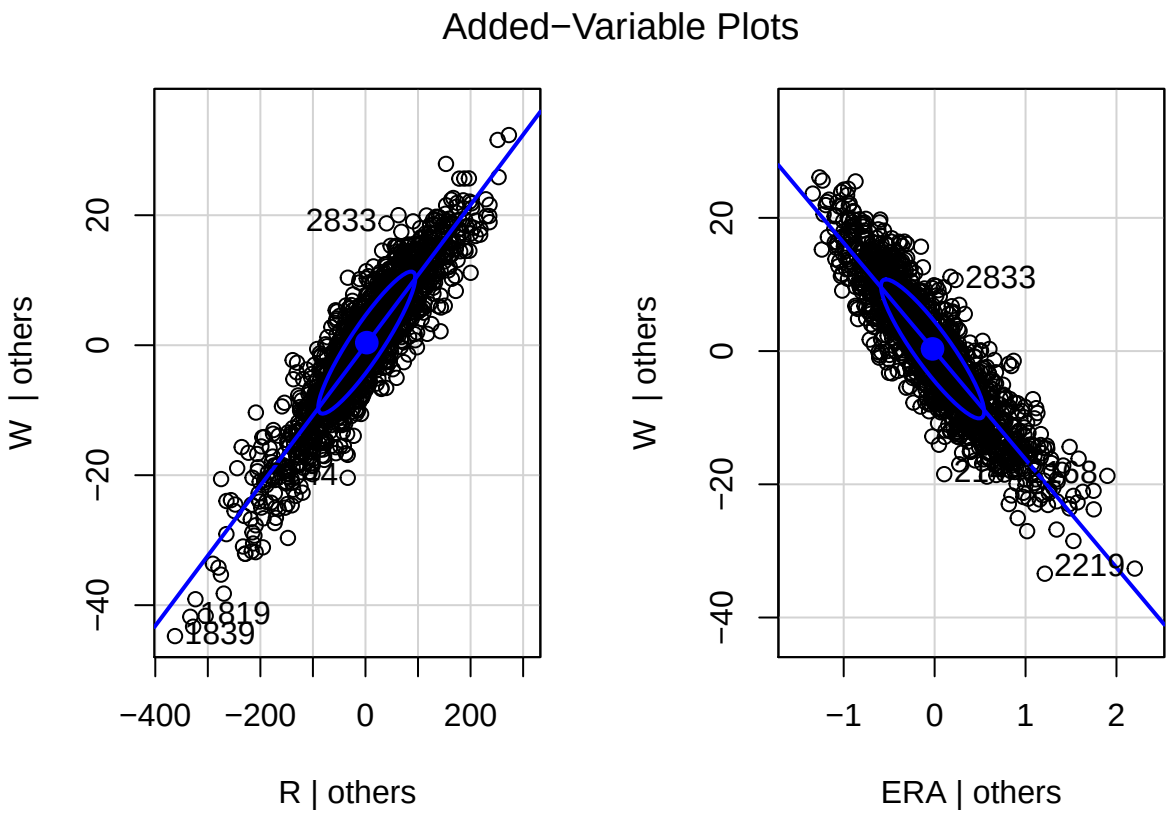
R^2: 87%, R^2_{ADJ}: 87%

These all indicate the variables are significant and this model is strong.

ANOVA:

term	df	sumsq	meansq	statistic	p.value
R	1	80080	80080	7732	1.321e-20
ERA	1	121071	121071	11690	7.345e-22
Residuals	1515	29490	19.47	NA	NA
Lack of fit	1501	29345	19.55	1.888	0.08305
Pure Error	14	145	10.36	NA	NA

IV. Results

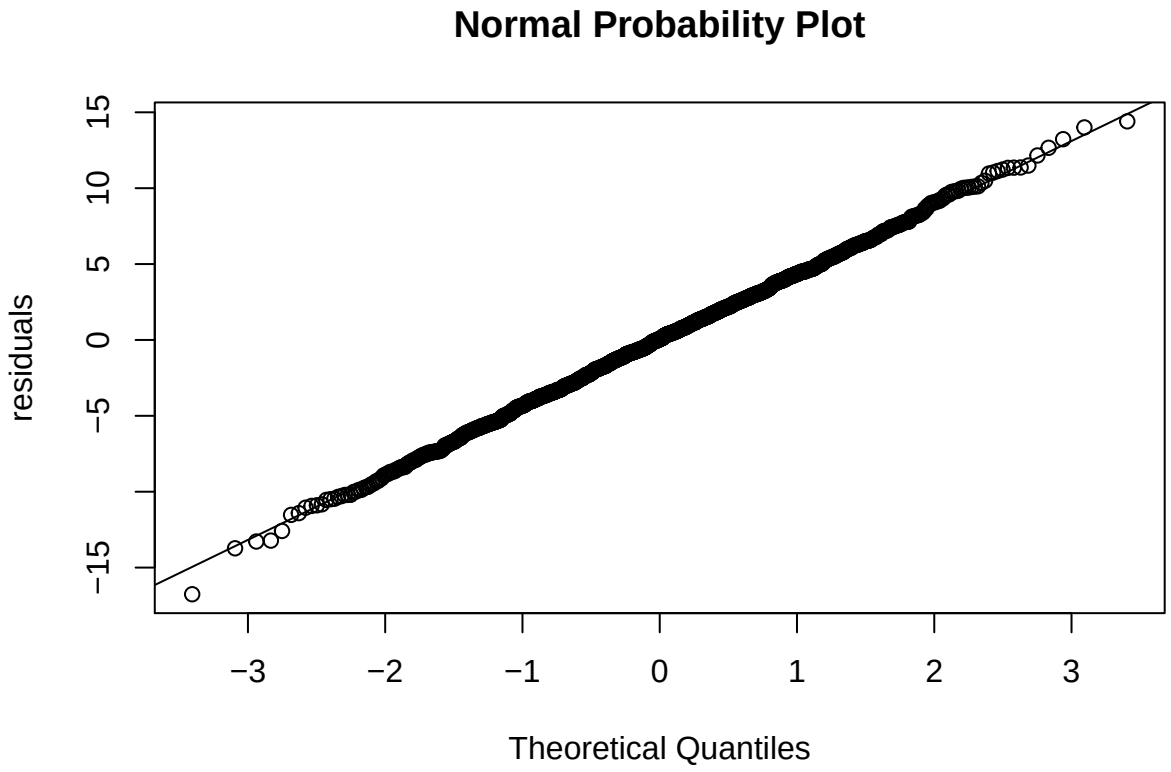


Both variables are linear. Runs are positive and ERA is negative. Both variables add value to the model.

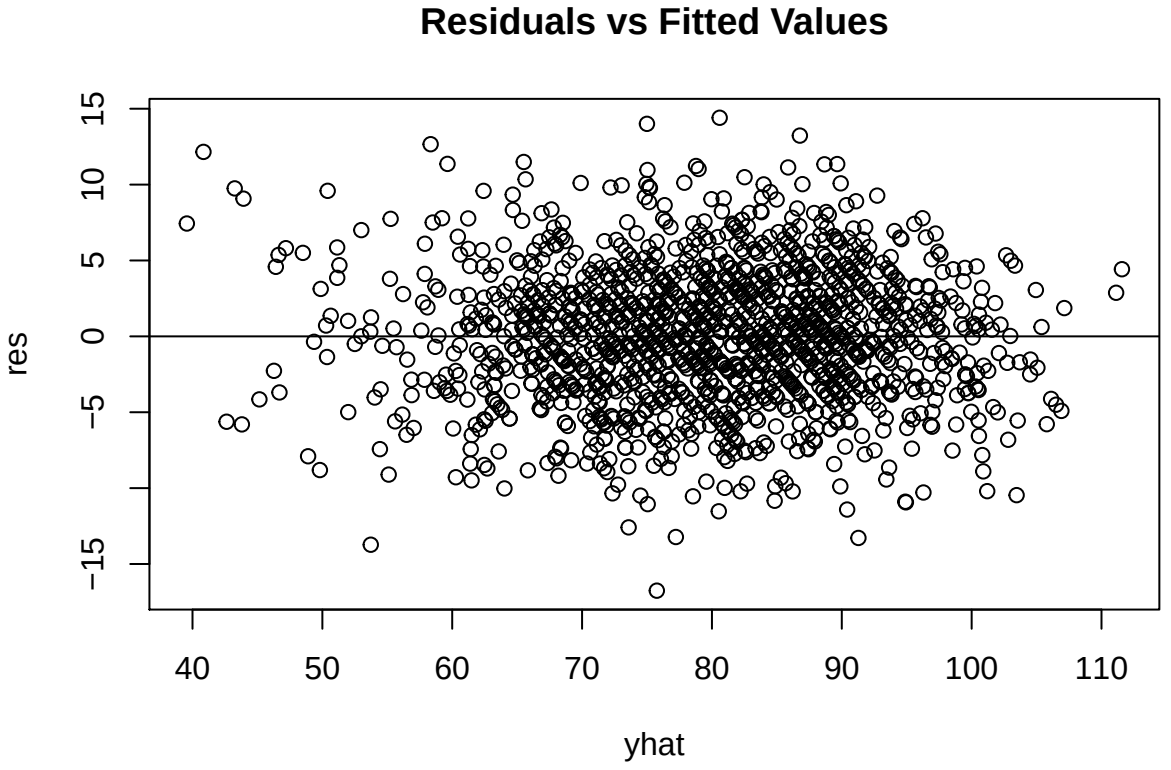
T-Distribution % Confidence Interval

```
## [1] "T = 1.9615"
```

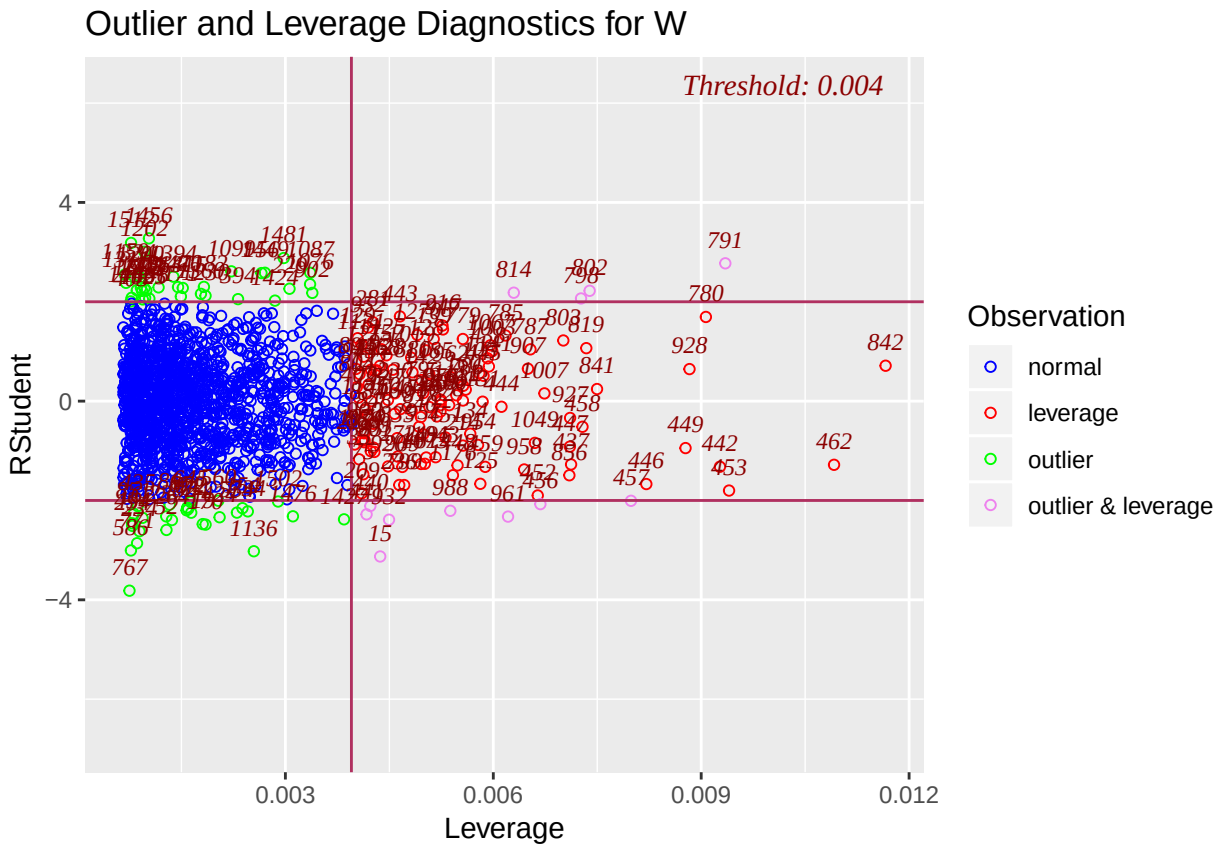
	2.5 %	97.5 %
(Intercept)	67.12	70.92
R	0.1056	0.1104
ERA	-16.66	-15.85



The residuals are linear.



The data might be slightly tailed to the right, but overall it looks to be distributed properly.

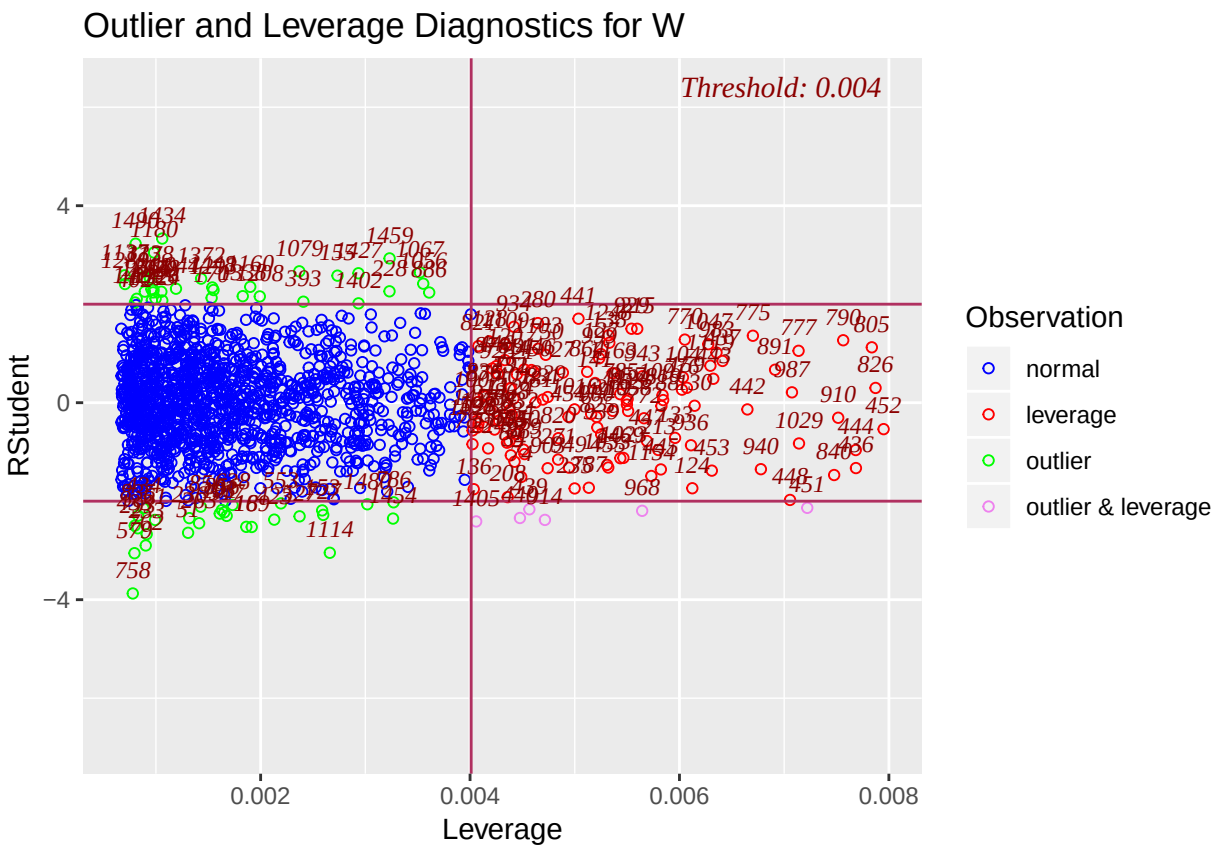


It looks like a few data points are outliers and leveraging the model. The next step is to remove those points.

```
teams.subset <- teams.subset[-c(15,968,961,457,453,462,842,928,780,791,798,814,802,911,
                                826,760,1116,758,1115,446,449,442,462),]
wins.lm2 <- lm(W ~ R + ERA, data = teams.subset)
summary(wins.lm2)
```

```
##
## Call:
## lm(formula = W ~ R + ERA, data = teams.subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.772  -2.926   0.074   2.898  14.447
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.663439   0.980124   71.08   <2e-16 ***
## R             0.107718   0.001261   85.45   <2e-16 ***
## ERA         -16.365050   0.209210  -78.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.352 on 1493 degrees of freedom
## Multiple R-squared:  0.8685, Adjusted R-squared:  0.8683
## F-statistic:  4928 on 2 and 1493 DF,  p-value: < 2.2e-16
```

Re-run plot with points removed.



Check for Multicollinearity

```
##
## Call:
## omcdiag(x = X, y = teams.subset$W)
##
##
## Overall Multicollinearity Diagnostics
##
##                        MC Results detection
## Determinant |X'X|:         0.8671         0
## Farrar Chi-Square:       212.5731         1
## Red Indicator:             0.3646         0
## Sum of Lambda Inverse:     2.3066         0
## Theil's Method:           -0.6026         0
## Condition Number:         18.4949         0
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test

##
## Call:
## imcdiag(x = X, y = teams.subset$W)
##
```

```
##
## All Individual Multicollinearity Diagnostics Result
##
##        VIF    TOL       Wi  Fi Leamer   CVIF Klein
## R   1.1533 0.8671 229.0151 Inf 0.9312 0.3405     0
## ERA 1.1533 0.8671 229.0151 Inf 0.9312 0.3405     0
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## * all coefficients have significant t-ratios
##
## R-square of y on all x: 0.8685
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

No collinearity detected from either test.

This model looks good to predict wins from an estimated runs total and ERA for the season. The few points we removed were unexpected, but removing them from the model appears to have strengthen it. Overall it appears that the linear model is the correct choice of a model.

V. Conclusion

This model looks to be reliable to predict MLB wins from projected runs scored and ERA. I think more variables would be good to help predict wins, but would I would worry about overfitting. Alot of baseball statistics are dependent on other statistics. ERA seems to have a large effect on win totals, which makes sense, the less runs you allow the more likely it is you will win.

VI. References

Data Collected from: SeanLahman.com

VII. Appendices

Dataset and code can be found on my GitHub: GitHub