# Baseball

STAT 4230 Group Project

James MacPherson, Eileen Shengaout, and Heather Cranford

# Background: “Moneyball”

- During the early 2000s, two workers for Oakland Athletics needed to recruit the best players for their team.
- They were on a budget, so they used statistical analysis to determine the best players for their team.
- The data set for this project comes from the same data set used by the workers.
- For our project, we utilized many of the variables from this data to predict

the amount of games won

Reference:
The data set for this project was uploaded by “WESDUCKET” seven years ago on the website, “Kaggle.” The data set is titled “Moneyball: MLB Statistics 1962-2012”

# Variables in the Baseball Data Set

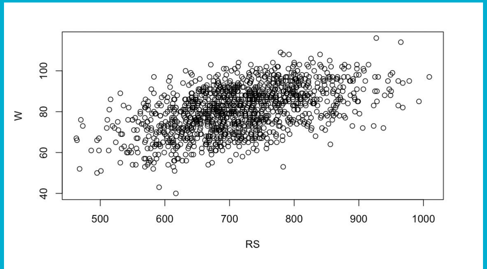
- Runs Scored in Total Over Course of Season (“RS”)
- Runs Allowed in Total Over Course of Season (“RA”)
- On Base Percentage (“OBP”)
- Slugging Percentage (“SLG”)
- Batting Average (“BA”)
- Playoffs (“Playoffs”) <- (Made Playoffs (“1”) or Missed Playoffs (“2”))
- Total Games in Season (“G”)
- Wins in Season (“W”)

# Scatterplots of the Dependent Variable Versus Each Variable

Simple Linear Regression

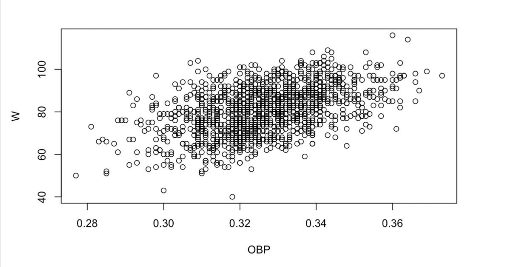
# Scatterplots

## (i) Wins Versus Runs Scored



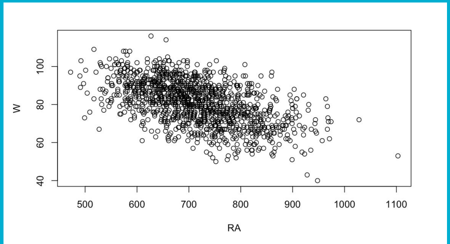
As the amount of runs scored increases, the amount of wins tend to increase.

## (iii) Wins Versus On Base Percentage



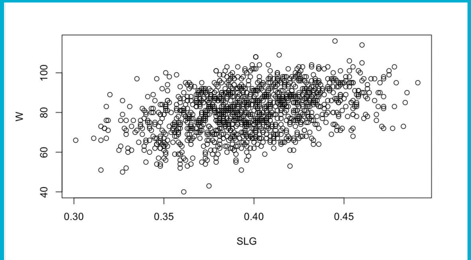
As the base percentage increases, the amount of wins tend to increase.

## (ii) Wins Versus Runs Allowed



As the amount of runs allowed increases, the amount of wins tend to decrease.

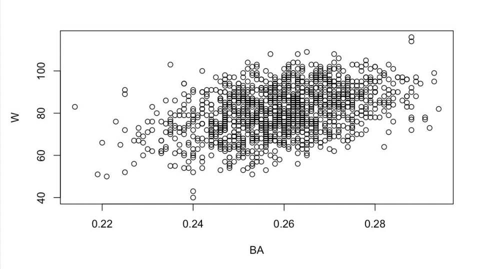
## (iv) Wins Versus Slugging



As slugging increases, the amount of wins tend to increase.

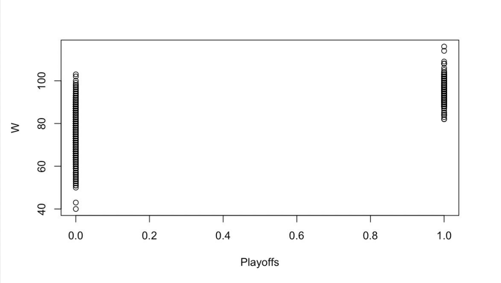
# Scatterplots

## (v) Wins Versus Batting Average

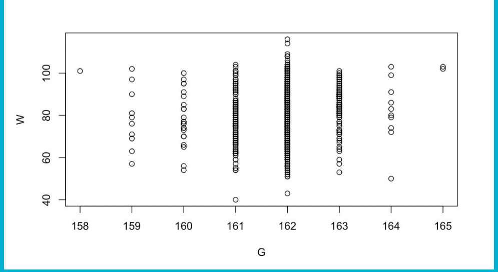


As the batting average increases, the amount of wins tend to increase.

## (vi) Wins Versus Playoffs

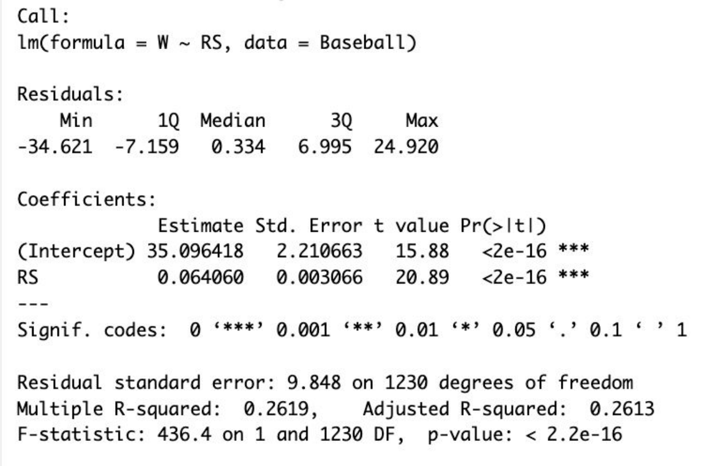


## (vii) Wins Versus Games



# **Simple Linear Regression:** Runs Scored in Total Over Course of Season (“RS”)

## Model Summary:



```
Call:
lm(formula = W ~ RS, data = Baseball)

Residuals:
    Min      1Q  Median      3Q     Max 
-34.621  -7.159   0.334   6.995  24.920 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 35.096418   2.210663   15.88   <2e-16 ***
RS           0.064060   0.003066   20.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.848 on 1230 degrees of freedom
Multiple R-squared:  0.2619,	Adjusted R-squared:  0.2613 
F-statistic: 436.4 on 1 and 1230 DF,  p-value: < 2.2e-16
```

## Assumptions:

- Linearity: satisfied
- Constant Variance: satisfied
- Normality: satisfied
- Independence: not satisfied

## Interpretation of Model:

Intercept: 35.096418

RS Slope: 0.064060

R^2: 0.2619

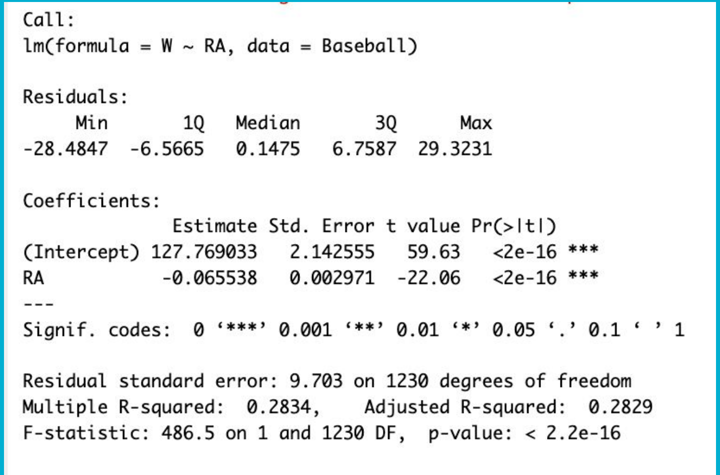
As the runs scored in total over the course of the season increases by one unit, the number of wins is predicted to increase by 0.064060.

26.19% of the variation in the number of wins is explained by the runs scored in total over course of season.

Predicted Number of Wins = 35.096418 + 0.064060(RS)

# Simple Linear Regression: Runs Allowed in Total Over Course of Season (“RA”)

## Model Summary:



```
Call:
lm(formula = W ~ RA, data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max 
-28.4847  -6.5665   0.1475   6.7587  29.3231 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 127.769033   2.142555   59.63   <2e-16 ***
RA           -0.065538   0.002971  -22.06   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.703 on 1230 degrees of freedom
Multiple R-squared:  0.2834,	Adjusted R-squared:  0.2829 
F-statistic: 486.5 on 1 and 1230 DF,  p-value: < 2.2e-16
```

## Assumptions:

- Linearity: satisfied
- Constant Variance: satisfied
- Normality: satisfied
- Independence: not satisfied

## Interpretation of Model:

Intercept: 127.769033

RS Slope: - 0.065538

R^2: 0.2834

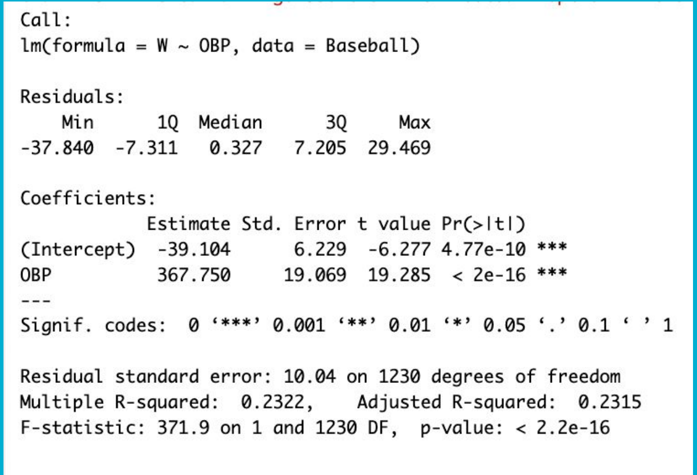
As the runs allowed in total over course of season increases by one unit, the number of wins is predicted to decrease by 0.065538.

28.34% of the variation in the number of wins is explained by runs allowed in total over course of season.

Predicted Number of Wins = 127.769033 - 0.065538(RA)

# Simple Linear Regression: On Base Percentage (“OBP”)

## Model Summary:



```
Call:
lm(formula = W ~ OBP, data = Baseball)

Residuals:
    Min      1Q  Median      3Q     Max
-37.840  -7.311   0.327   7.205  29.469

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -39.104      6.229  -6.277 4.77e-10 ***
OBP          367.750     19.069  19.285  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.04 on 1230 degrees of freedom
Multiple R-squared:  0.2322,	Adjusted R-squared:  0.2315
F-statistic: 371.9 on 1 and 1230 DF,  p-value: < 2.2e-16
```

## Assumptions:

- Linearity: satisfied
- Constant Variance: not satisfied
- Normality: not satisfied
- Independence: satisfied

## Interpretation of Model:

Intercept: -39.104

RS Slope: 367.750

R^2: 0.2322

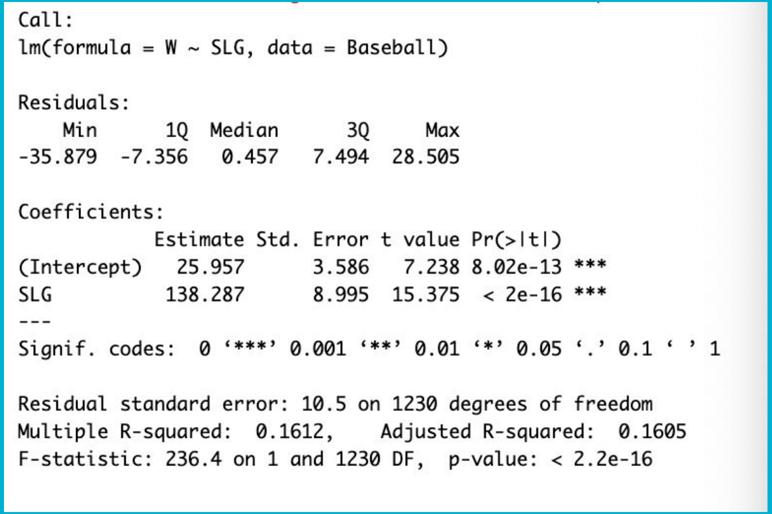
As the on base percentage increases by one unit, the number of wins is predicted to increase by 367.750.

23.22% of the variation in the number of wins is explained by the on base percentage.

Predicted Number of Wins = -39.104 + 367.750(OBP)

# Simple Linear Regression: Slugging Percentage (“SLG”)

## Model Summary:



```
Call:
lm(formula = W ~ SLG, data = Baseball)

Residuals:
    Min      1Q  Median      3Q     Max 
-35.879  -7.356   0.457   7.494  28.505 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   25.957      3.586   7.238 8.02e-13 ***
SLG          138.287      8.995  15.375  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.5 on 1230 degrees of freedom
Multiple R-squared:  0.1612,	Adjusted R-squared:  0.1605 
F-statistic: 236.4 on 1 and 1230 DF,  p-value: < 2.2e-16
```

## Assumptions:

- Linearity: satisfied
- Constant Variance: satisfied
- Normality: not satisfied
- Independence: satisfied

## Interpretation of Model:

Intercept: 25.957

RS Slope: 138.287

R^2: 0.1612

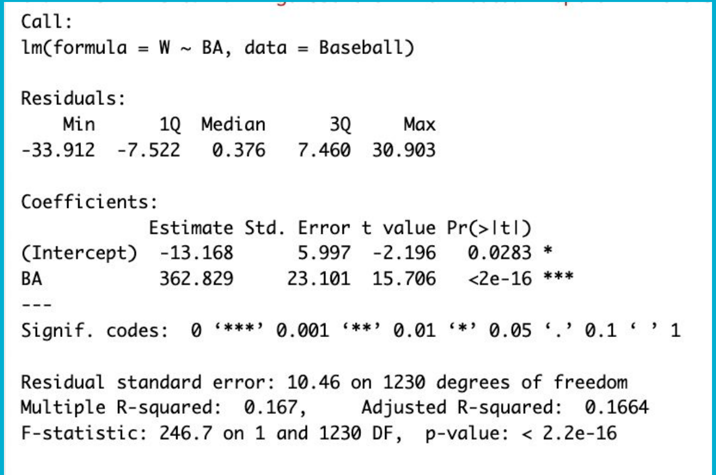
As the slugging percentage increases by one unit, the number of wins is predicted to increase by 138.287.

16.12% of the variation in the number of wins is explained by the slugging percentage.

Predicted Number of Wins = 25.957 + 138.287(SLG)

# Simple Linear Regression: Batting Average (“BA”)

## Model Summary:



```
Call:
lm(formula = W ~ BA, data = Baseball)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.912  -7.522   0.376   7.460  30.903 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -13.168      5.997  -2.196   0.0283 *  
BA           362.829     23.101  15.706   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.46 on 1230 degrees of freedom
Multiple R-squared:  0.167,	Adjusted R-squared:  0.1664 
F-statistic: 246.7 on 1 and 1230 DF,  p-value: < 2.2e-16
```

## Assumptions:

- Linearity: satisfied
- Constant Variance: not satisfied
- Normality: not satisfied
- Independence: satisfied

## Interpretation of Model:

Intercept: -13.168

RS Slope: 362.829

R^2: 0.167

As the batting average increases by one unit, the number of wins is predicted to increase by 362.829.

16.7% of the variation in the number of wins is explained by the batting average.

Predicted Number of Wins = -13.168 + 362.829(BA)

# Summary of Simple Linear Regression

- In order to predict the number of wins, a simple linear regression model will not be the best representation.

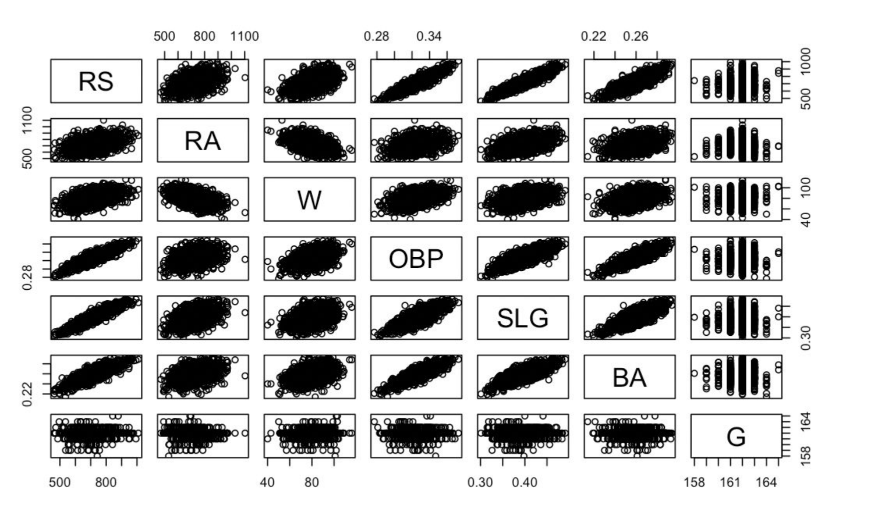
    - Each model has at least one assumption that is violated.

    - Although each model using one predictor has a p-value of less than our alpha of 0.5, the R^2 values are not large enough to be accepted as the best model for predicting wins.

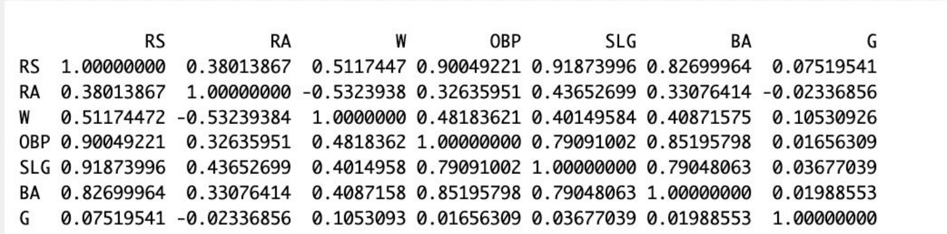
# Scatterplot Matrix of the data

# Multiple Linear Regression

# Relationship Between the Number of Wins and Each of the Predictors



## Correlation Coefficients:



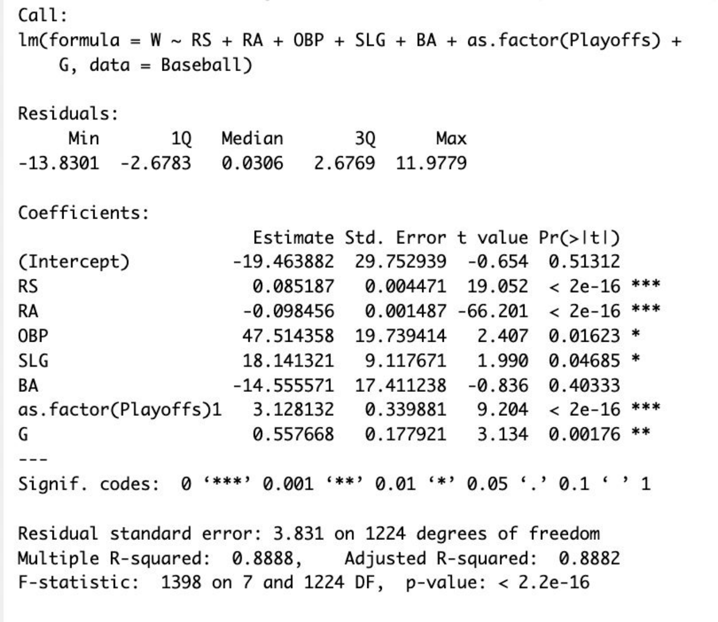
```
            RS          RA          W        OBP        SLG         BA           G
RS  1.00000000  0.38013867  0.5117447 0.90049221 0.91873996 0.82699964  0.07519541
RA  0.38013867  1.00000000 -0.5323938 0.32635951 0.43652699 0.33076414 -0.02336856
W   0.51174472 -0.53239384  1.0000000 0.48183621 0.40149584 0.40871575  0.10530926
OBP 0.90049221  0.32635951  0.4818362 1.00000000 0.79091002 0.85195798  0.01656309
SLG 0.91873996  0.43652699  0.4014958 0.79091002 1.00000000 0.79048063  0.03677039
BA  0.82699964  0.33076414  0.4087158 0.85195798 0.79048063 1.00000000  0.01988553
G   0.07519541 -0.02336856  0.1053093 0.01656309 0.03677039 0.01988553  1.00000000
```

This information displays the relationship the variables.

There are some variables that show a linear trend, but this is expected. In baseball, a team that obtains a higher volume of wins would have "good statistics" in many different categories.
Every variable shows somewhat of a linear trend when compared to "W" (Wins in a Season). When looking at multicollinearity, we will need to take which response variables are highly correlated.

# Multiple Linear Regression

## Model Summary Output: Using all 7 Predictors



```
Call:
lm(formula = W ~ RS + RA + OBP + SLG + BA + as.factor(Playoffs) +
    G, data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8301  -2.6783   0.0306   2.6769  11.9779

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -19.463882  29.752939  -0.654  0.51312
RS                     0.085187   0.004471  19.052  < 2e-16 ***
RA                    -0.098456   0.001487 -66.201  < 2e-16 ***
OBP                   47.514358  19.739414   2.407  0.01623 *
SLG                   18.141321   9.117671   1.990  0.04685 *
BA                   -14.555571  17.411238  -0.836  0.40333
as.factor(Playoffs)1   3.128132   0.339881   9.204  < 2e-16 ***
G                      0.557668   0.177921   3.134  0.00176 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.831 on 1224 degrees of freedom
Multiple R-squared:  0.8888,    Adjusted R-squared:  0.8882
F-statistic:  1398 on 7 and 1224 DF,  p-value: < 2.2e-16
```

Predicted Number of Wins = -19.463882 + 0.085187(RS) - 0.098456(RA) + 47.514358(OBP) + 18.141321(SLG) - 14.555571(BA) + 3.128132(Playoffs) + 0.557668(G)

# Interpretations of Multiple Linear Regression Model

Slope for RS: As the runs scored in total over the course of the season increases by one unit, the number of wins is predicted to increase by 0.085187 while holding the other predictors in the model fixed.

Slope for RA:  As the runs allowed in total over course of season increases by one unit, the number of wins is predicted to decrease by 0.098456 while holding the other predictors in the model fixed.

Slope for OBP: As the on base percentage increases by one unit,  the number of wins is predicted to increase by 47.514358 while holding the other predictors in the model fixed.

Slope for SLG: As the slugging percentage increases by one unit, the number of wins is predicted to increase by 18.141321 while holding the other predictors in the model fixed.

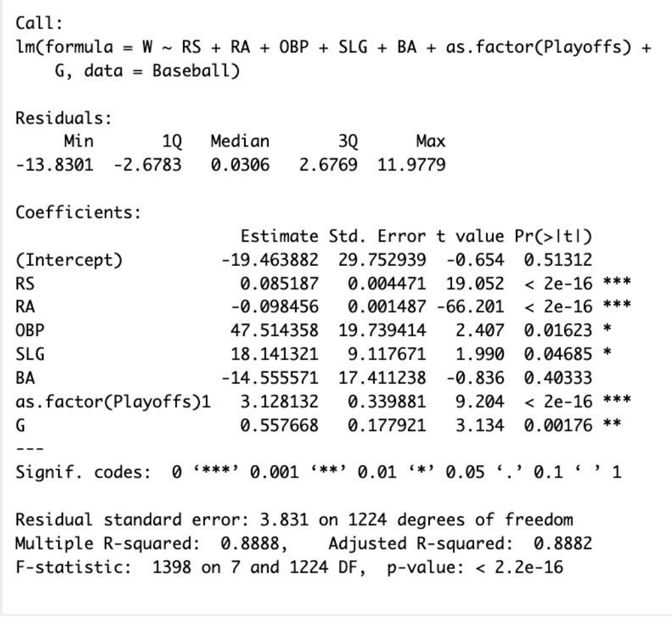
# Interpretations of Multiple Linear Regression Model (continued)

Slope for BA: As the batting average increases by one unit, the number of wins is predicted to decrease by 14.555571 while holding the other predictors in the model fixed.

Slope for Playoffs: If a team makes the playoffs (1), the predicted wins is expected to increase by 3.128132, keeping all other variables the same.

Slope for G: As the number of games increases by one unit, the number of wins is predicted to increase by 0.557668 while holding the other predictors in the model fixed.

# Residuals: Typical Size of Prediction Errors

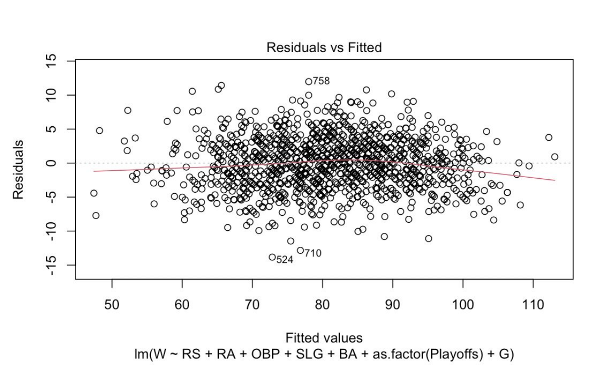


```
Call:
lm(formula = W ~ RS + RA + OBP + SLG + BA + as.factor(Playoffs) +
    G, data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8301  -2.6783   0.0306   2.6769  11.9779

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -19.463882  29.752939  -0.654  0.51312
RS                     0.085187   0.004471  19.052  < 2e-16 ***
RA                    -0.098456   0.001487 -66.201  < 2e-16 ***
OBP                   47.514358  19.739414   2.407  0.01623 *
SLG                   18.141321   9.117671   1.990  0.04685 *
BA                   -14.555571  17.411238  -0.836  0.40333
as.factor(Playoffs)1   3.128132   0.339881   9.204  < 2e-16 ***
G                      0.557668   0.177921   3.134  0.00176 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.831 on 1224 degrees of freedom
Multiple R-squared:  0.8888,    Adjusted R-squared:  0.8882
F-statistic:  1398 on 7 and 1224 DF,  p-value: < 2.2e-16
```

Using this model with all seven predictors for predicting the number of wins will predict the number of wins with an average error of 3.831, also known as the Residual Standard Error.

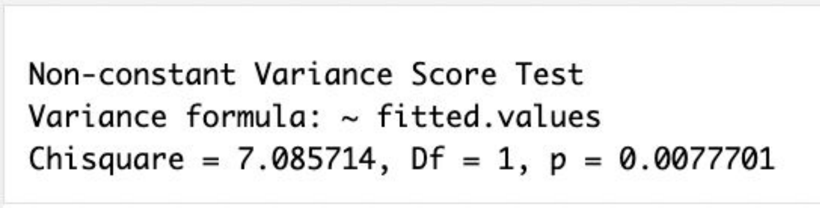
# Checking Assumptions for the Multiple Linear Regression Model

## Linearity:



There is a linear pattern surrounding the horizontal line at 0. Therefore, the linearity condition is satisfied for this model.

## Constant Variance:

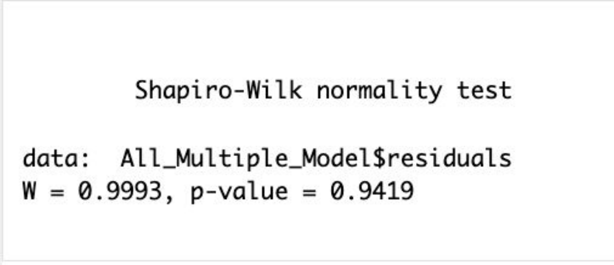


```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 7.085714, Df = 1, p = 0.0077701
```

Using the Breusch Pagan Test:
We rejected the null hypothesis (residuals have constant variance) and concluded
that the residuals do not have constant variance.

# Checking Assumptions for the Multiple Linear Regression Model (continued)

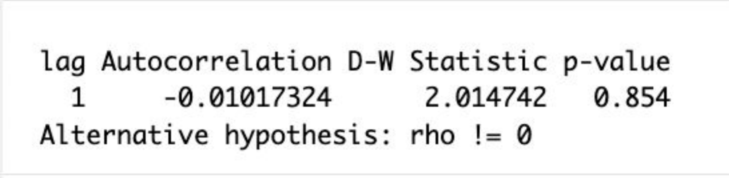
## Normality:



```
        Shapiro-Wilk normality test

data:  All_Multiple_Model$residuals
W = 0.9993, p-value = 0.9419
```

Using the Shapiro Wilks Test:
We failed to reject the null hypothesis (residuals are normally distributed) and concluded that the residuals are normally distributed.

## Independence:

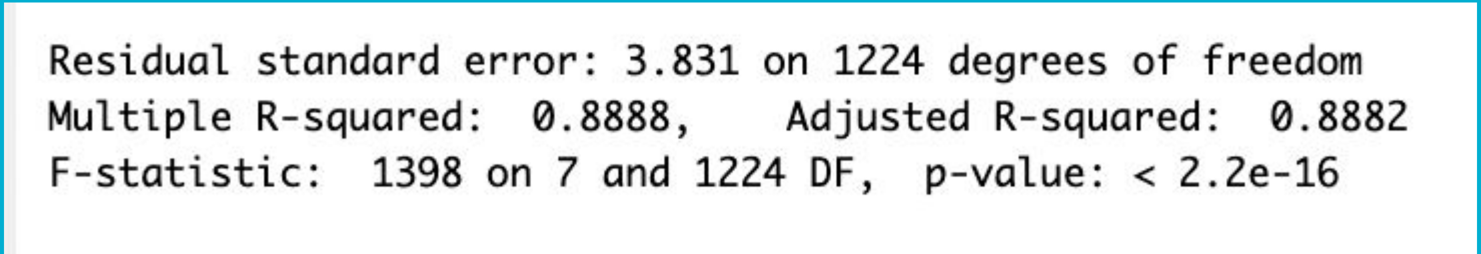


```
 lag Autocorrelation D-W Statistic p-value
   1     -0.01017324      2.014742   0.854
 Alternative hypothesis: rho != 0
```

Using the Durbin Watson Test:
We failed to reject the null hypothesis (residuals are independent) and concluded the residuals are independent.

In summary, 3 of 4 assumptions have been satisfied for the multiple linear regression model.

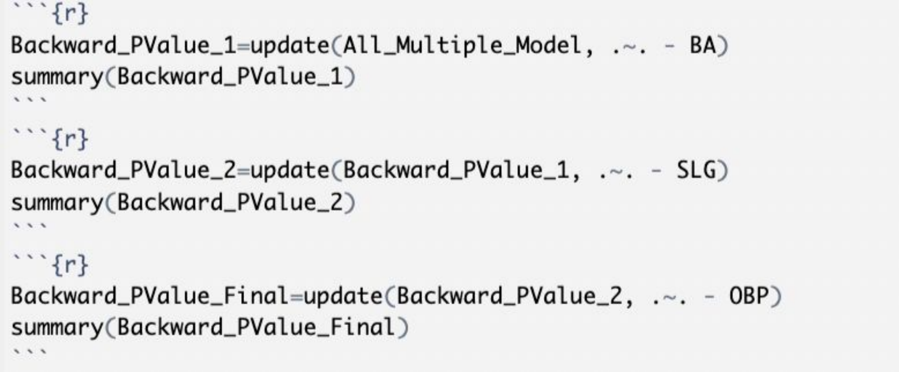
# Interpretation of R^2 Value



```
Residual standard error: 3.831 on 1224 degrees of freedom
Multiple R-squared:  0.8888,	Adjusted R-squared:  0.8882
F-statistic:  1398 on 7 and 1224 DF,  p-value: < 2.2e-16
```

The R^2 value is 0.8888, so 88.88% of the variation in the number of wins is explained by this model using all seven predictors.

# Selection of Variables: Backward Elimination with P-value as Criteria



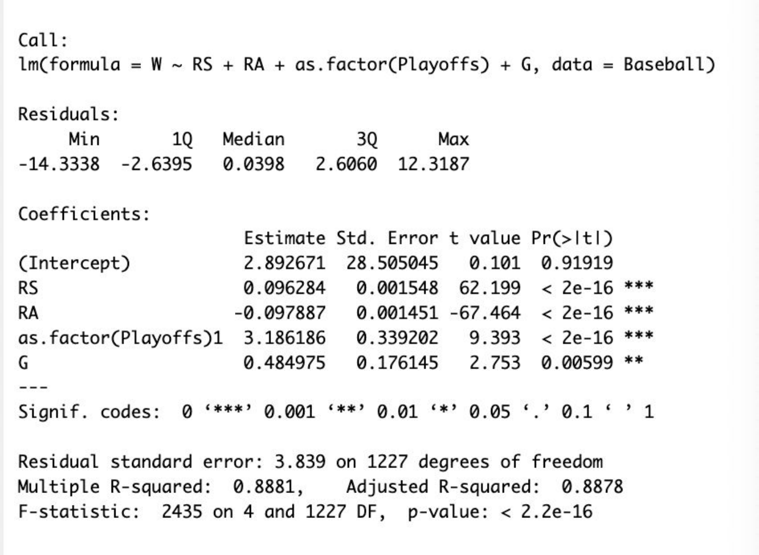
~~~
```{r}
Backward_PValue_1=update(All_Multiple_Model, .~. - BA)
summary(Backward_PValue_1)
```
```{r}
Backward_PValue_2=update(Backward_PValue_1, .~. - SLG)
summary(Backward_PValue_2)
```
```{r}
Backward_PValue_Final=update(Backward_PValue_2, .~. - OBP)
summary(Backward_PValue_Final)
```
~~~

Using this method, predictors were eliminated one at a time based off of their p-value.

This elimination left us with four predictors for the multiple linear regression model to represent this data set.

# Backward Elimination with P-value as Criteria

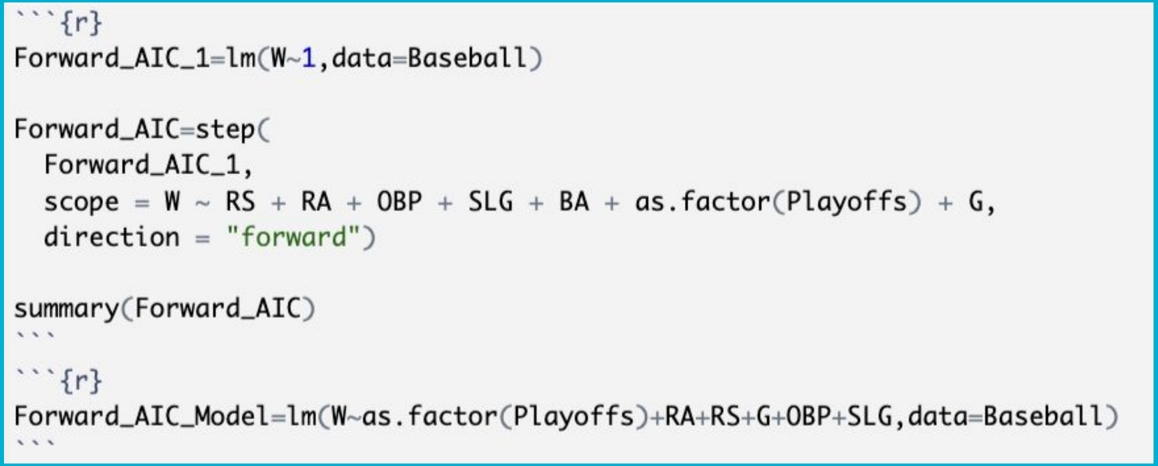
## Output Summary of Final Model Using the Method:



```
Call:
lm(formula = W ~ RS + RA + as.factor(Playoffs) + G, data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.3338  -2.6395   0.0398   2.6060  12.3187 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           2.892671  28.505045   0.101  0.91919    
RS                    0.096284   0.001548  62.199  < 2e-16 ***
RA                   -0.097887   0.001451 -67.464  < 2e-16 ***
as.factor(Playoffs)1  3.186186   0.339202   9.393  < 2e-16 ***
G                     0.484975   0.176145   2.753  0.00599 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.839 on 1227 degrees of freedom
Multiple R-squared:  0.8881,	Adjusted R-squared:  0.8878 
F-statistic:  2435 on 4 and 1227 DF,  p-value: < 2.2e-16
```

By the end of this method, the variables’ p-values are less than an alpha of 0.05. Therefore, the variables that should be included in the model based off of backward elimination are RS, RA, Playoffs, and G.

# Selection of Variables: Forward Selection with AIC as Criteria

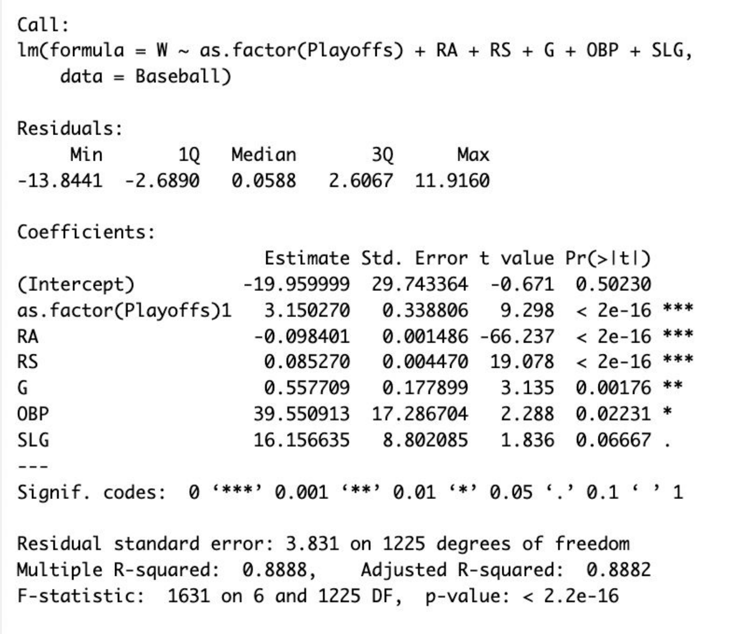


~~~
```{r}
Forward_AIC_1=lm(W~1,data=Baseball)

Forward_AIC=step(
  Forward_AIC_1,
  scope = W ~ RS + RA + OBP + SLG + BA + as.factor(Playoffs) + G,
  direction = "forward")

summary(Forward_AIC)
~~~

```{r}
Forward_AIC_Model=lm(W~as.factor(Playoffs)+RA+RS+G+OBP+SLG,data=Baseball)
```
```

Forward AIC was performed in order to determine which variables would best fit our model. This process was carried out until the lowest AIC value possible was achieved.
```

# Forward Selection with AIC as Criteria

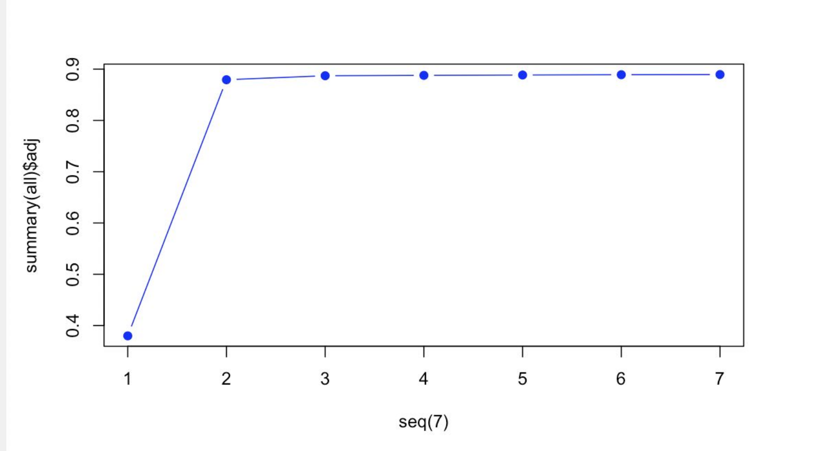
## Output Summary of Final Model Using the Method:



```
Call:
lm(formula = W ~ as.factor(Playoffs) + RA + RS + G + OBP + SLG,
    data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.8441  -2.6890   0.0588   2.6067  11.9160 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -19.959999  29.743364  -0.671  0.50230    
as.factor(Playoffs)1   3.150270   0.338806   9.298  < 2e-16 ***
RA                    -0.098401   0.001486 -66.237  < 2e-16 ***
RS                     0.085270   0.004470  19.078  < 2e-16 ***
G                      0.557709   0.177899   3.135  0.00176 ** 
OBP                   39.550913  17.286704   2.288  0.02231 *  
SLG                   16.156635   8.802085   1.836  0.06667 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.831 on 1225 degrees of freedom
Multiple R-squared:  0.8888,	Adjusted R-squared:  0.8882 
F-statistic:  1631 on 6 and 1225 DF,  p-value: < 2.2e-16
```

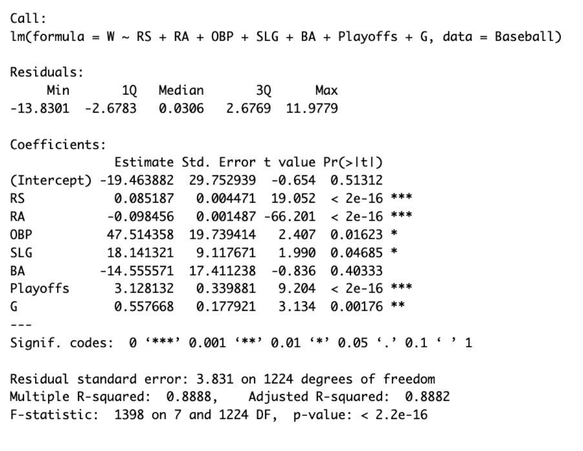
By the end of this method, six predictors were left in our model. The predictors that should be included in the model based off of backward elimination are Playoffs, RA, RS, G, OBP, and SLG.

# Variable Selection: R-Squared Adjusted Criteria



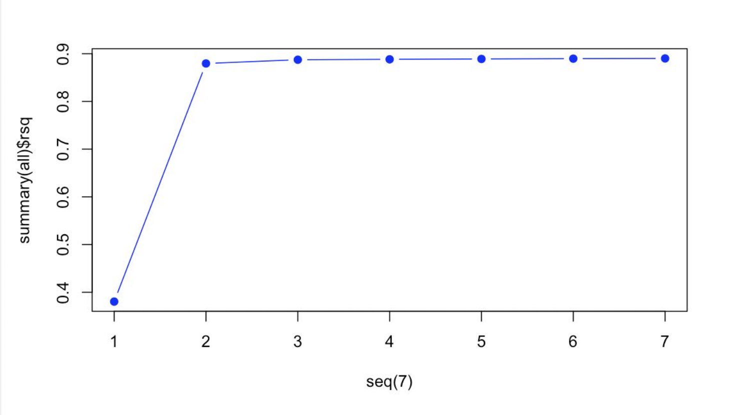
R^2 decreases when a predictor that is not useful is added to the model, so, based on this criteria, all seven predictors should be used in the model.

Summary Output



```
Call:
lm(formula = W ~ RS + RA + OBP + SLG + BA + Playoffs + G, data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.8301  -2.6783   0.0306   2.6769  11.9779 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -19.463882  29.752939  -0.654  0.51312    
RS            0.085187   0.004471  19.052  < 2e-16 ***
RA           -0.098456   0.001487 -66.201  < 2e-16 ***
OBP          47.514358  19.739414   2.407  0.01623 *  
SLG          18.141321   9.117671   1.990  0.04685 *  
BA          -14.555571  17.411238  -0.836  0.40333    
Playoffs      3.128132   0.339881   9.204  < 2e-16 ***
G             0.557668   0.177921   3.134  0.00176 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.831 on 1224 degrees of freedom
Multiple R-squared:  0.8888,    Adjusted R-squared:  0.8882 
F-statistic:  1398 on 7 and 1224 DF,  p-value: < 2.2e-16
```

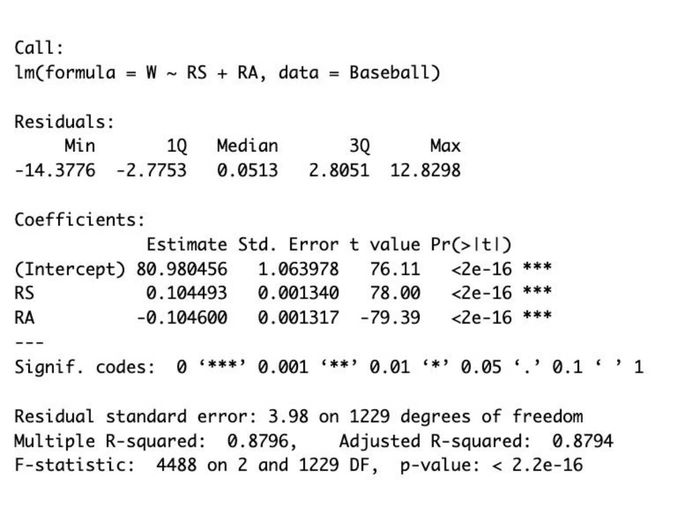
# Variable Selection: R-Squared Criteria



According to the R-Squared Criteria, the R-Square stops increasing dramatically when there are two predictors in the model. These two predictors are RS and RA.

# R-Squared Criteria

Output Summary of Final Model Using the Method:

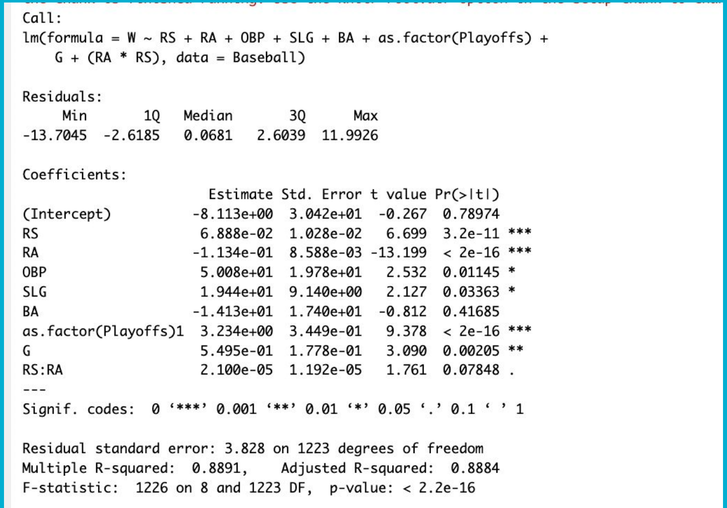


```
Call:
lm(formula = W ~ RS + RA, data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.3776  -2.7753   0.0513   2.8051  12.8298 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 80.980456   1.063978   76.11   <2e-16 ***
RS           0.104493   0.001340   78.00   <2e-16 ***
RA          -0.104600   0.001317  -79.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.98 on 1229 degrees of freedom
Multiple R-squared:  0.8796,	Adjusted R-squared:  0.8794 
F-statistic:  4488 on 2 and 1229 DF,  p-value: < 2.2e-16
```

The predictors we should use in our model based off the R-Squared Criteria are RS and RA.

# Interaction Model

- We tested if RS and RA interacted.
- First, we tested whether the overall model was useful with the interaction term.

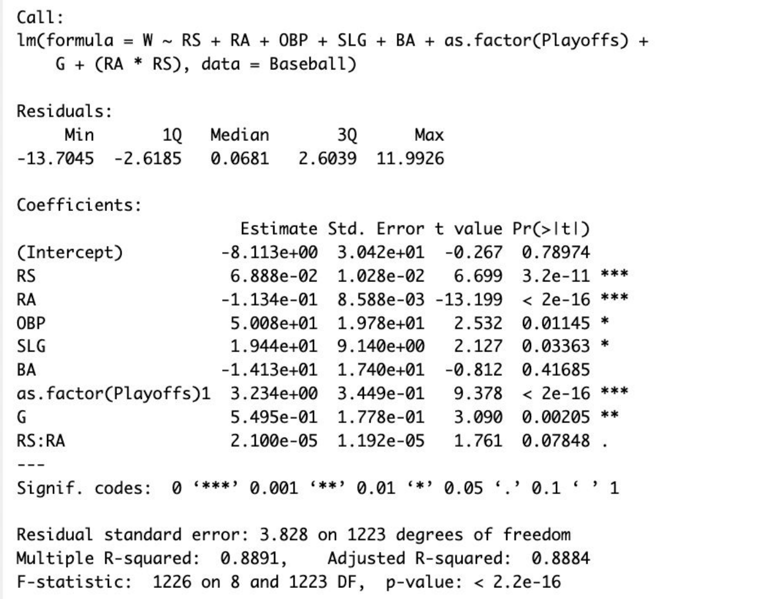


```
Call:
lm(formula = W ~ RS + RA + OBP + SLG + BA + as.factor(Playoffs) +
    G + (RA * RS), data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max
-13.7045  -2.6185   0.0681   2.6039  11.9926

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -8.113e+00  3.042e+01  -0.267  0.78974
RS                    6.888e-02  1.028e-02   6.699  3.2e-11 ***
RA                   -1.134e-01  8.588e-03 -13.199  < 2e-16 ***
OBP                   5.008e+01  1.978e+01   2.532  0.01145 *
SLG                   1.944e+01  9.140e+00   2.127  0.03363 *
BA                   -1.413e+01  1.740e+01  -0.812  0.41685
as.factor(Playoffs)1  3.234e+00  3.449e-01   9.378  < 2e-16 ***
G                     5.495e-01  1.778e-01   3.090  0.00205 **
RS:RA                 2.100e-05  1.192e-05   1.761  0.07848 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.828 on 1223 degrees of freedom
Multiple R-squared:  0.8891,    Adjusted R-squared:  0.8884
F-statistic:  1226 on 8 and 1223 DF,  p-value: < 2.2e-16
```

Since the p-value is <2.2e^-16, we can reject the null hypothesis (the overall model is not useful) and conclude that overall model is useful.

# Interaction Model

- Then, we tested to see if the interaction term is useful.



```
Call:
lm(formula = W ~ RS + RA + OBP + SLG + BA + as.factor(Playoffs) +
    G + (RA * RS), data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max
-13.7045  -2.6185   0.0681   2.6039  11.9926

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -8.113e+00  3.042e+01  -0.267  0.78974
RS                   6.888e-02  1.028e-02   6.699  3.2e-11 ***
RA                  -1.134e-01  8.588e-03 -13.199  < 2e-16 ***
OBP                  5.008e+01  1.978e+01   2.532  0.01145 *
SLG                  1.944e+01  9.140e+00   2.127  0.03363 *
BA                  -1.413e+01  1.740e+01  -0.812  0.41685
as.factor(Playoffs)1  3.234e+00  3.449e-01   9.378  < 2e-16 ***
G                    5.495e-01  1.778e-01   3.090  0.00205 **
RS:RA                2.100e-05  1.192e-05   1.761  0.07848 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.828 on 1223 degrees of freedom
Multiple R-squared:  0.8891,	Adjusted R-squared:  0.8884
F-statistic:  1226 on 8 and 1223 DF,  p-value: < 2.2e-16
```

Since the p-value for the interaction term is 0.07848 which is greater than alpha of 0.05, we fail to reject the null hypothesis (the interaction is not useful) and conclude that the interaction between RS and RA is not useful for our model.

# Quadratic Model

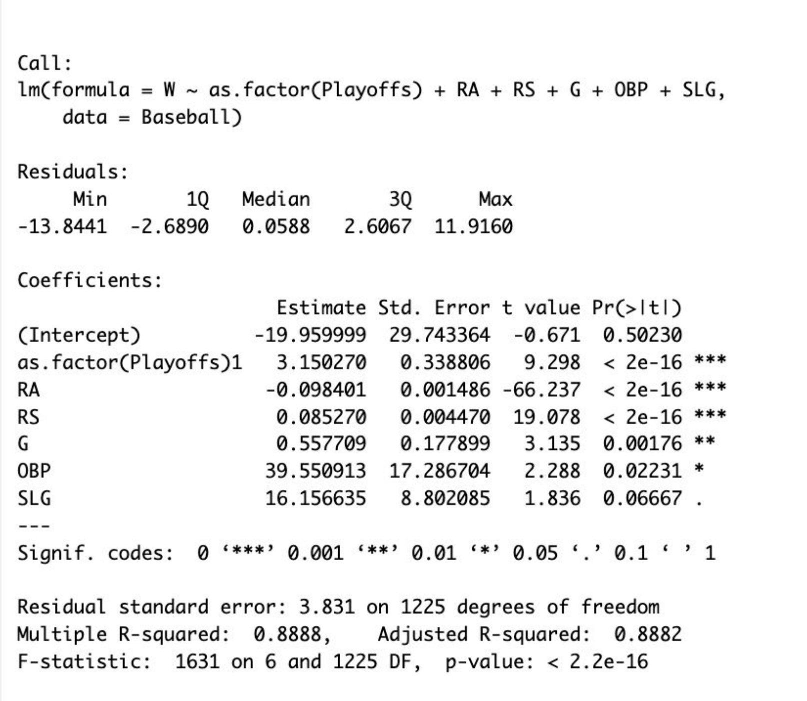
Since all of our quantitative variables have linear trends, we concluded that it would not be appropriate to include quadratic terms for our best model.

Quadratic Terms would not accurately represent the baseball data set.

Since we are not including quadratic terms, there is not a necessity to run Partial Nested F-Tests.

# Final Two Models

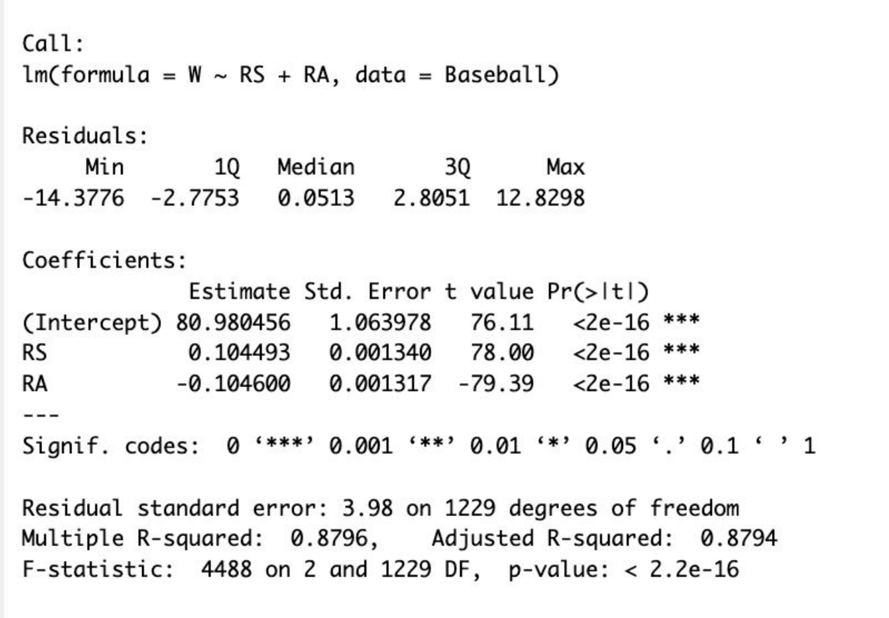
Forward_AIC_Model:



```
Call:
lm(formula = W ~ as.factor(Playoffs) + RA + RS + G + OBP + SLG,
    data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8441  -2.6890   0.0588   2.6067  11.9160

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -19.959999  29.743364  -0.671  0.50230
as.factor(Playoffs)1  3.150270   0.338806   9.298  < 2e-16 ***
RA                   -0.098401   0.001486 -66.237  < 2e-16 ***
RS                    0.085270   0.004470  19.078  < 2e-16 ***
G                     0.557709   0.177899   3.135  0.00176 **
OBP                  39.550913  17.286704   2.288  0.02231 *
SLG                  16.156635   8.802085   1.836  0.06667 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.831 on 1225 degrees of freedom
Multiple R-squared:  0.8888,    Adjusted R-squared:  0.8882
F-statistic:  1631 on 6 and 1225 DF,  p-value: < 2.2e-16
```

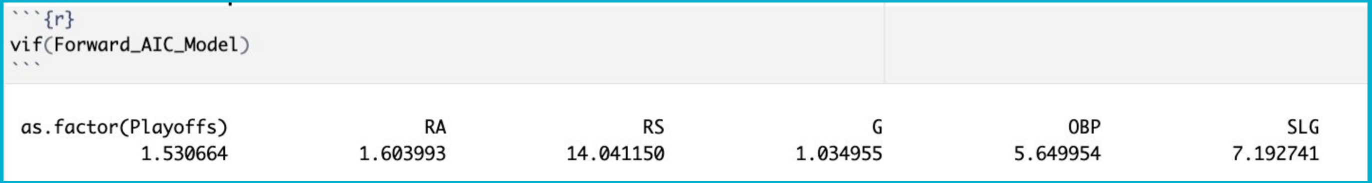
AND

RSquare_Model:



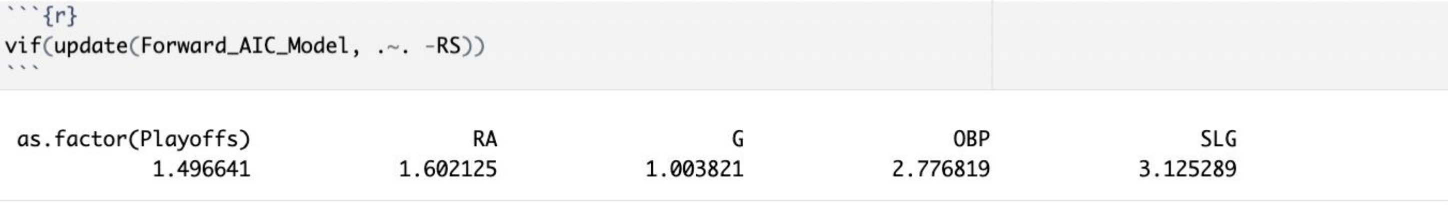
```
Call:
lm(formula = W ~ RS + RA, data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max
-14.3776  -2.7753   0.0513   2.8051  12.8298

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 80.980456   1.063978   76.11   <2e-16 ***
RS           0.104493   0.001340   78.00   <2e-16 ***
RA          -0.104600   0.001317  -79.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.98 on 1229 degrees of freedom
Multiple R-squared:  0.8796,    Adjusted R-squared:  0.8794
F-statistic:  4488 on 2 and 1229 DF,  p-value: < 2.2e-16
```

# Multicollinearity: Model 1 (Forward_AIC_Model)



~~~
```{r}
vif(Forward_AIC_Model)
```

as.factor(Playoffs)                  RA                  RS                   G                 OBP                 SLG 
           1.530664            1.603993           14.041150            1.034955            5.649954            7.192741 
~~~

The VIF value for RS is greater than 10, so it will be taken out of the model.

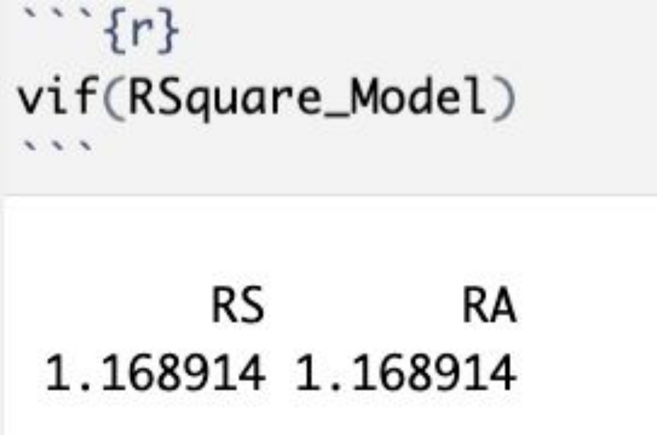


~~~
```{r}
vif(update(Forward_AIC_Model, .~. -RS))
```

as.factor(Playoffs)                  RA                   G                 OBP                 SLG 
           1.496641            1.602125            1.003821            2.776819            3.125289 
~~~

All predictors have a VIF value of less than 10.

The final model will include Playoffs, RA, G, OBP, and SLG.

# Multicollinearity: Model 2 (RSqaure_Model)

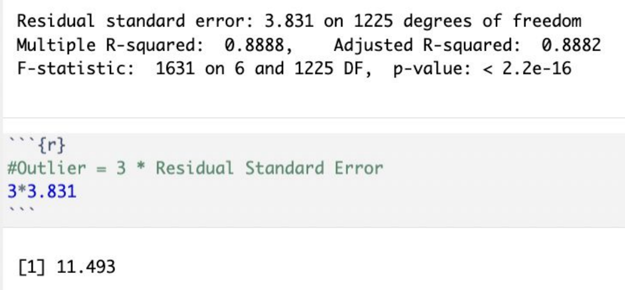


~~~
```{r}
vif(RSquare_Model)
```
~~~

```
      RS       RA 
1.168914 1.168914 
```

Both predictors have a VIF value of less than 10, so they will both be included in the final model.

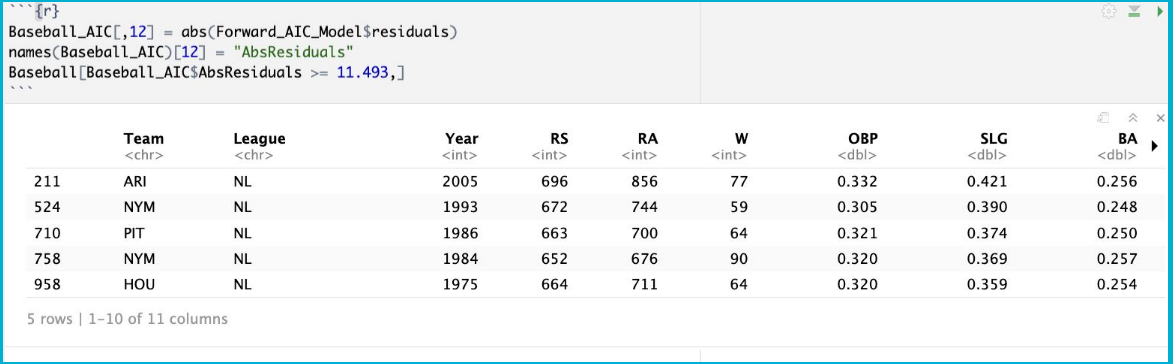
# Checking for Outliers: Model 1 (Forward_AIC_Model)



```
Residual standard error: 3.831 on 1225 degrees of freedom
Multiple R-squared:  0.8888,	Adjusted R-squared:  0.8882 
F-statistic:  1631 on 6 and 1225 DF,  p-value: < 2.2e-16
```

~~~
```{r}
#Outlier = 3 * Residual Standard Error
3*3.831
```
~~~

```
[1] 11.493
```



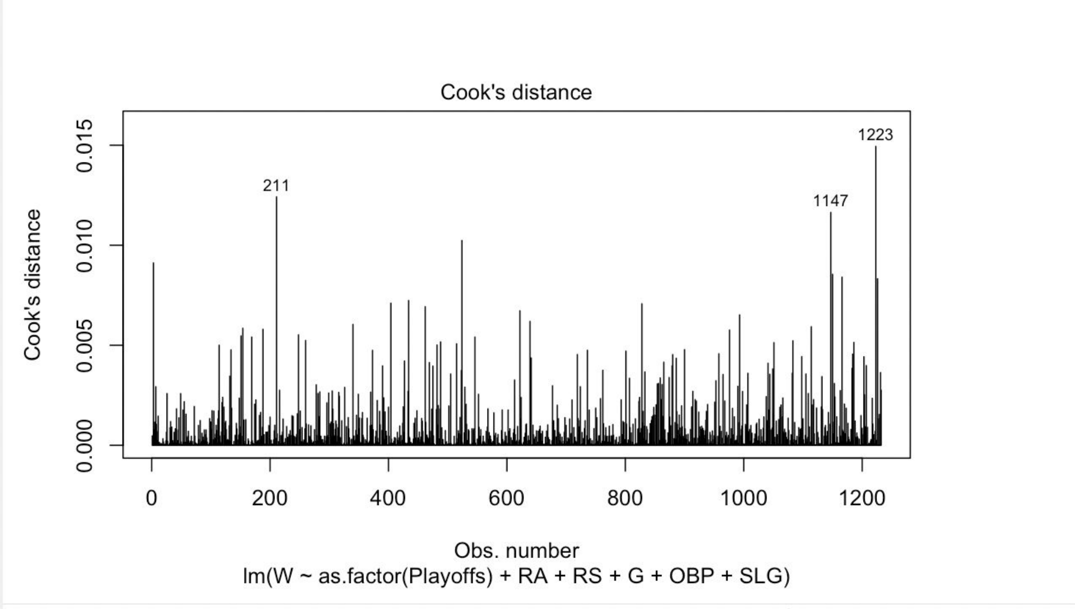
~~~
```{r}
Baseball_AIC[,12] = abs(Forward_AIC_Model$residuals)
names(Baseball_AIC)[12] = "AbsResiduals"
Baseball[Baseball_AIC$AbsResiduals >= 11.493,]
```
~~~

| | Team<br><chr> | League<br><chr> | Year<br><int> | RS<br><int> | RA<br><int> | W<br><int> | OBP<br><dbl> | SLG<br><dbl> | BA<br><dbl> |
|---|---|---|---|---|---|---|---|---|---|
| 211 | ARI | NL | 2005 | 696 | 856 | 77 | 0.332 | 0.421 | 0.256 |
| 524 | NYM | NL | 1993 | 672 | 744 | 59 | 0.305 | 0.390 | 0.248 |
| 710 | PIT | NL | 1986 | 663 | 700 | 64 | 0.321 | 0.374 | 0.250 |
| 758 | NYM | NL | 1984 | 652 | 676 | 90 | 0.320 | 0.369 | 0.257 |
| 958 | HOU | NL | 1975 | 664 | 711 | 64 | 0.320 | 0.359 | 0.254 |

5 rows | 1-10 of 11 columns

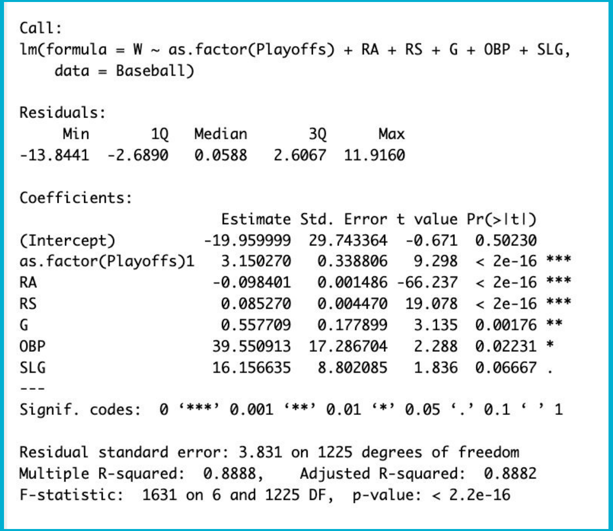
The values with absolute residuals greater than 11.493 will be taken out of the data since they are outliers.

# Cook’s Distance to find Highly Influential Points Model 1 (Forward_AIC_Model)



# Highly Influential Points

Once the highly influential points were determined, the data set was updated excluding these points. The new data is called Baseball_AIC_Data.



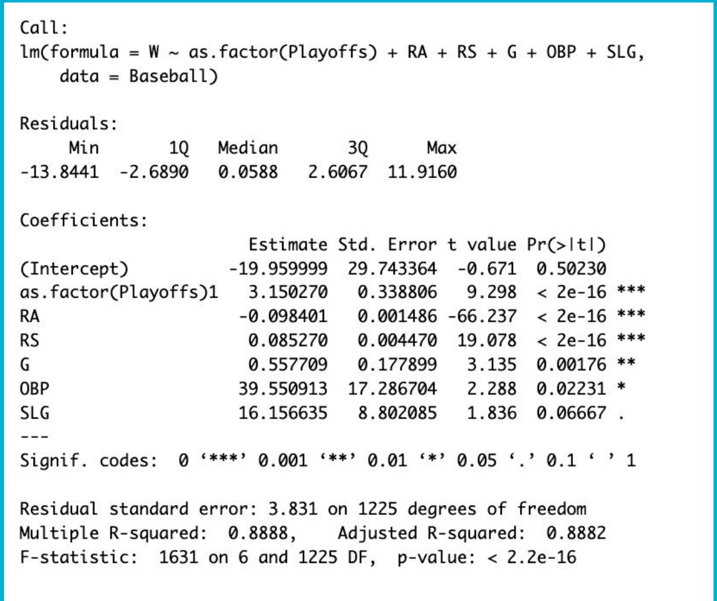
```
Call:
lm(formula = W ~ as.factor(Playoffs) + RA + RS + G + OBP + SLG,
    data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.8441  -2.6890   0.0588   2.6067  11.9160 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -19.959999  29.743364  -0.671  0.50230    
as.factor(Playoffs)1   3.150270   0.338806   9.298  < 2e-16 ***
RA                    -0.098401   0.001486 -66.237  < 2e-16 ***
RS                     0.085270   0.004470  19.078  < 2e-16 ***
G                      0.557709   0.177899   3.135  0.00176 ** 
OBP                   39.550913  17.286704   2.288  0.02231 *  
SLG                   16.156635   8.802085   1.836  0.06667 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.831 on 1225 degrees of freedom
Multiple R-squared:  0.8888,	Adjusted R-squared:  0.8882 
F-statistic:  1631 on 6 and 1225 DF,  p-value: < 2.2e-16
```

A new model (AIC_Model2) was made with this new data.

# Comparing R^2 Values

The R^2 value of the new model was then compared with the Forward_AIC_Model.

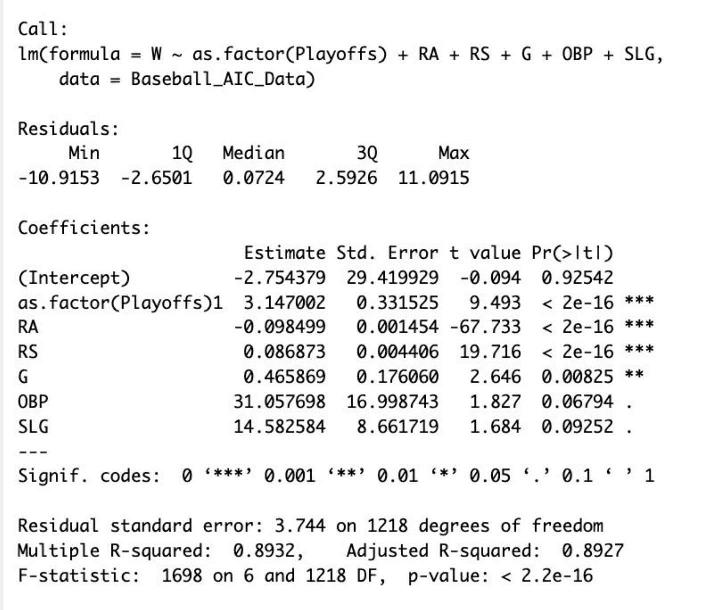
Forward_AIC_Model: R^2= 88.88%



```
Call:
lm(formula = W ~ as.factor(Playoffs) + RA + RS + G + OBP + SLG,
    data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8441  -2.6890   0.0588   2.6067  11.9160

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -19.959999  29.743364  -0.671  0.50230
as.factor(Playoffs)1  3.150270   0.338806   9.298  < 2e-16 ***
RA                   -0.098401   0.001486 -66.237  < 2e-16 ***
RS                    0.085270   0.004470  19.078  < 2e-16 ***
G                     0.557709   0.177899   3.135  0.00176 **
OBP                  39.550913  17.286704   2.288  0.02231 *
SLG                  16.156635   8.802085   1.836  0.06667 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.831 on 1225 degrees of freedom
Multiple R-squared:  0.8888,	Adjusted R-squared:  0.8882
F-statistic:  1631 on 6 and 1225 DF,  p-value: < 2.2e-16
```

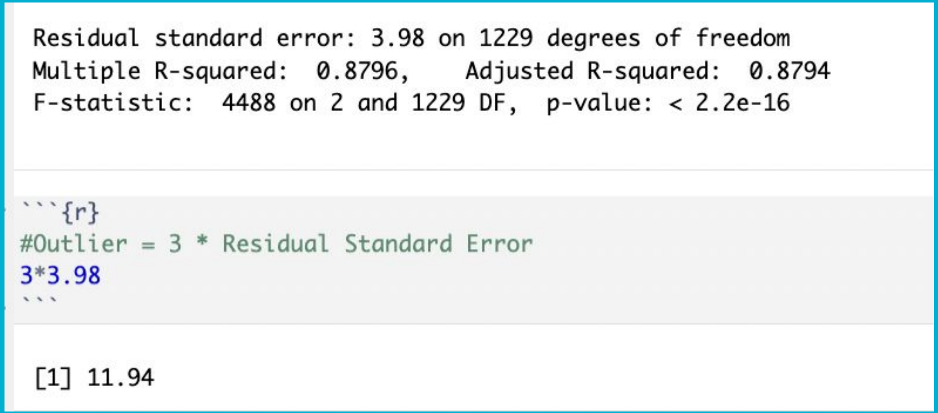
The model excluding the outliers and highly influential points produces a higher value of R^2 meaning that the variation in the number of wins is explained better by the AIC_Model2.

AIC_Model2: R^2 = 89.32%



```
Call:
lm(formula = W ~ as.factor(Playoffs) + RA + RS + G + OBP + SLG,
    data = Baseball_AIC_Data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9153  -2.6501   0.0724   2.5926  11.0915

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -2.754379  29.419929  -0.094  0.92542
as.factor(Playoffs)1  3.147002   0.331525   9.493  < 2e-16 ***
RA                   -0.098499   0.001454 -67.733  < 2e-16 ***
RS                    0.086873   0.004406  19.716  < 2e-16 ***
G                     0.465869   0.176060   2.646  0.00825 **
OBP                  31.057698  16.998743   1.827  0.06794 .
SLG                  14.582584   8.661719   1.684  0.09252 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.744 on 1218 degrees of freedom
Multiple R-squared:  0.8932,	Adjusted R-squared:  0.8927
F-statistic:  1698 on 6 and 1218 DF,  p-value: < 2.2e-16
```

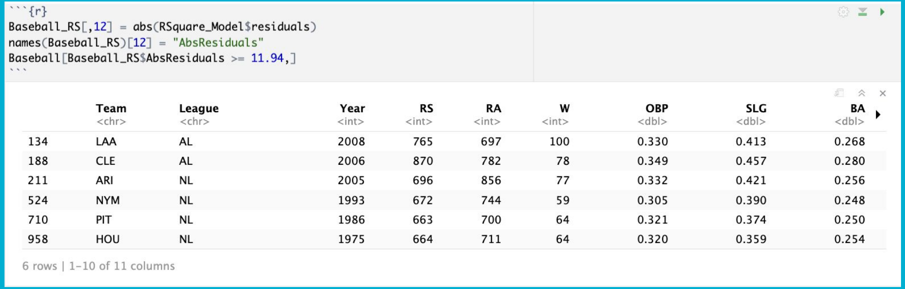
# Checking for Outliers: Model 2 (RSquare_Model)



```
Residual standard error: 3.98 on 1229 degrees of freedom
Multiple R-squared:  0.8796,    Adjusted R-squared:  0.8794
F-statistic:  4488 on 2 and 1229 DF,  p-value: < 2.2e-16
```

~~~
```{r}
#Outlier = 3 * Residual Standard Error
3*3.98
```
~~~

```
[1] 11.94
```



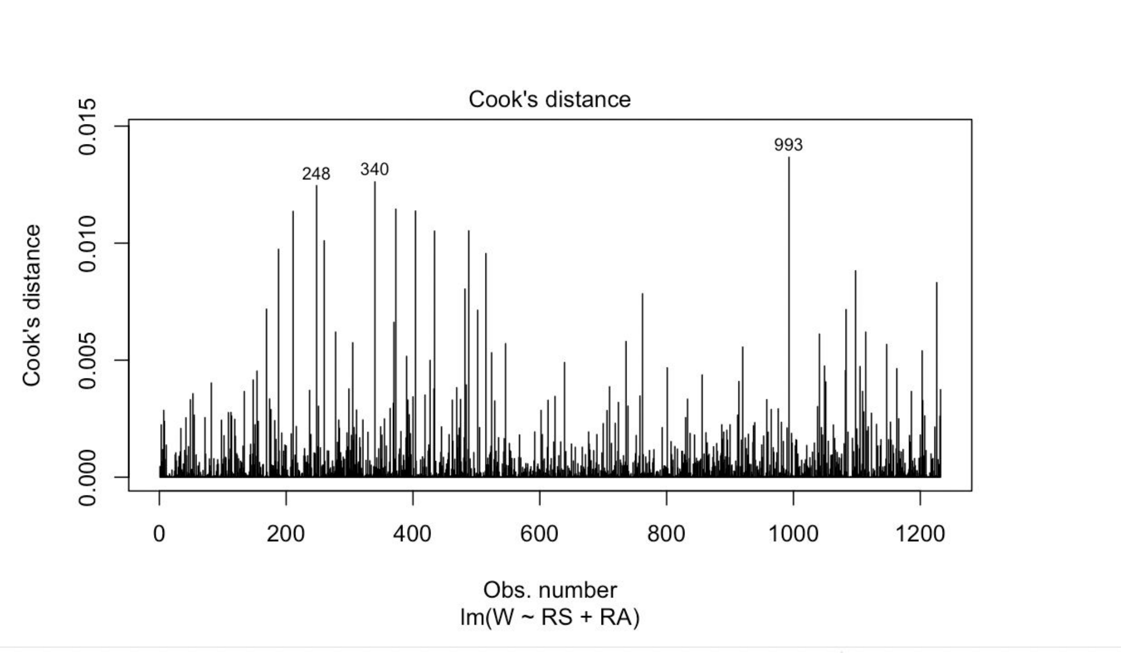
~~~
```{r}
Baseball_RS[,12] = abs(RSquare_Model$residuals)
names(Baseball_RS)[12] = "AbsResiduals"
Baseball[Baseball_RS$AbsResiduals >= 11.94,]
```
~~~

| | Team<br>&lt;chr&gt; | League<br>&lt;chr&gt; | Year<br>&lt;int&gt; | RS<br>&lt;int&gt; | RA<br>&lt;int&gt; | W<br>&lt;int&gt; | OBP<br>&lt;dbl&gt; | SLG<br>&lt;dbl&gt; | BA<br>&lt;dbl&gt; |
|---|---|---|---|---|---|---|---|---|---|
| 134 | LAA | AL | 2008 | 765 | 697 | 100 | 0.330 | 0.413 | 0.268 |
| 188 | CLE | AL | 2006 | 870 | 782 | 78 | 0.349 | 0.457 | 0.280 |
| 211 | ARI | NL | 2005 | 696 | 856 | 77 | 0.332 | 0.421 | 0.256 |
| 524 | NYM | NL | 1993 | 672 | 744 | 59 | 0.305 | 0.390 | 0.248 |
| 710 | PIT | NL | 1986 | 663 | 700 | 64 | 0.321 | 0.374 | 0.250 |
| 958 | HOU | NL | 1975 | 664 | 711 | 64 | 0.320 | 0.359 | 0.254 |

6 rows | 1-10 of 11 columns

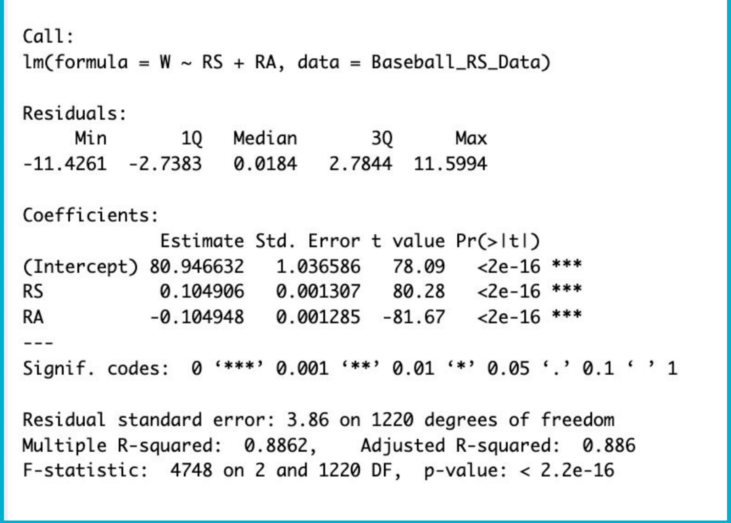
The values with absolute residuals greater than 11.94 will be taken out of the data since they are outliers.

# Cook's Distance to find Highly Influential Points Model 2 (RSquare_Model)



# Highly Influential Points

Once the highly influential points were determined, the data set was updated excluding these points. The new data is called Baseball_RS_Data.



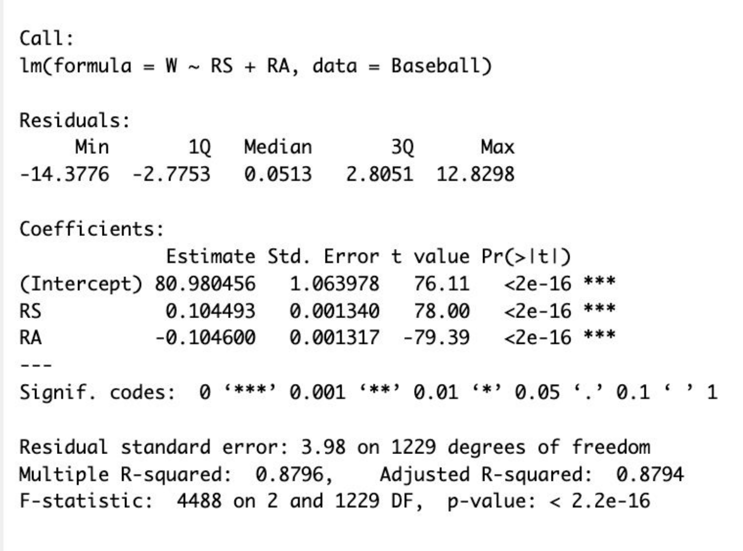
```
Call:
lm(formula = W ~ RS + RA, data = Baseball_RS_Data)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.4261  -2.7383   0.0184   2.7844  11.5994 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 80.946632   1.036586   78.09   <2e-16 ***
RS           0.104906   0.001307   80.28   <2e-16 ***
RA          -0.104948   0.001285  -81.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.86 on 1220 degrees of freedom
Multiple R-squared:  0.8862,	Adjusted R-squared:  0.886 
F-statistic:  4748 on 2 and 1220 DF,  p-value: < 2.2e-16
```

A new model (RS_Model2) was made with this new data.

# Comparing R^2 Values

The R^2 value of the new model was then compared with the RSquare_Model.

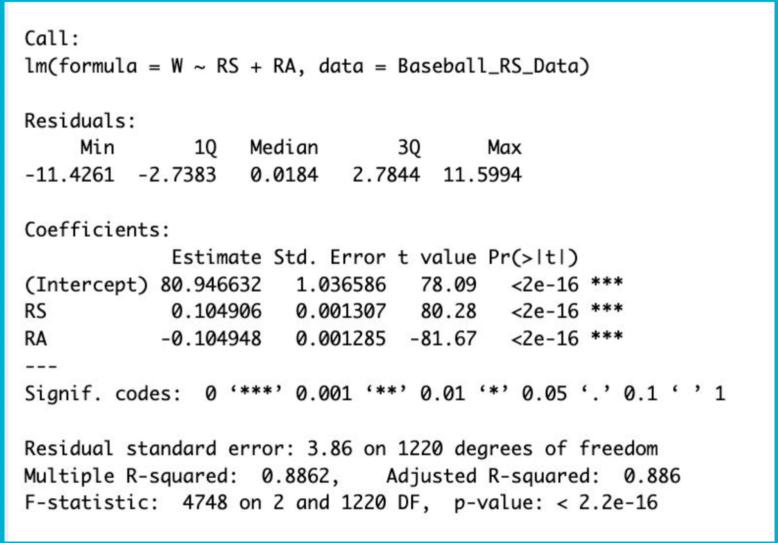
RSquare_Model = 87.96%



```
Call:
lm(formula = W ~ RS + RA, data = Baseball)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.3776  -2.7753   0.0513   2.8051  12.8298 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 80.980456   1.063978   76.11   <2e-16 ***
RS           0.104493   0.001340   78.00   <2e-16 ***
RA          -0.104600   0.001317  -79.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.98 on 1229 degrees of freedom
Multiple R-squared:  0.8796,	Adjusted R-squared:  0.8794 
F-statistic:  4488 on 2 and 1229 DF,  p-value: < 2.2e-16
```

The model excluding the outliers and highly influential points produces a higher value of R^2 meaning that the variation in the number of wins is explained better by the RS_Model2.

RS_Model2 = 88.62%

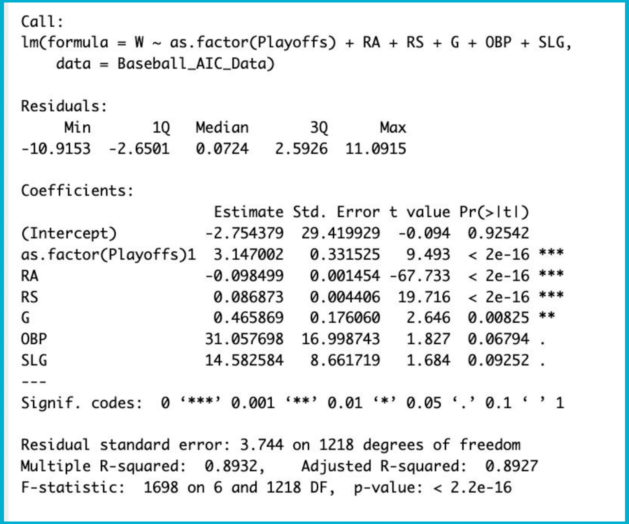


```
Call:
lm(formula = W ~ RS + RA, data = Baseball_RS_Data)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.4261  -2.7383   0.0184   2.7844  11.5994 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 80.946632   1.036586   78.09   <2e-16 ***
RS           0.104906   0.001307   80.28   <2e-16 ***
RA          -0.104948   0.001285  -81.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.86 on 1220 degrees of freedom
Multiple R-squared:  0.8862,	Adjusted R-squared:  0.886 
F-statistic:  4748 on 2 and 1220 DF,  p-value: < 2.2e-16
```

# The Final Model

The final model is the model, AIC_Model2 because this model produced the highest R^2 value of 89.32%.

- 89.32% of the variation in the number of wins is predicted by this model.



```
Call:
lm(formula = W ~ as.factor(Playoffs) + RA + RS + G + OBP + SLG,
    data = Baseball_AIC_Data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.9153  -2.6501   0.0724   2.5926  11.0915 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -2.754379  29.419929  -0.094  0.92542    
as.factor(Playoffs)1  3.147002   0.331525   9.493  < 2e-16 ***
RA                   -0.098499   0.001454 -67.733  < 2e-16 ***
RS                    0.086873   0.004406  19.716  < 2e-16 ***
G                     0.465869   0.176060   2.646  0.00825 ** 
OBP                  31.057698  16.998743   1.827  0.06794 .  
SLG                  14.582584   8.661719   1.684  0.09252 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.744 on 1218 degrees of freedom
Multiple R-squared:  0.8932,	Adjusted R-squared:  0.8927 
F-statistic:  1698 on 6 and 1218 DF,  p-value: < 2.2e-16
```

Predicted Number of Wins = -2.754379 + 3.127002(Playoffs) - 0.098499(RA) + 0.086873(RS) + 0.465869(G) + 31.057698(OBP) + 14.582584(SLG)

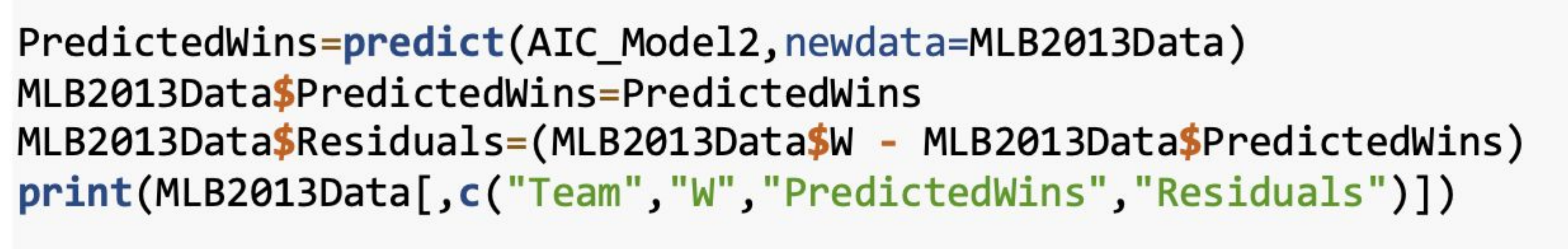
The predictors are:

- Playoffs
- RA
- RS
- G
- OBP
- SLG

P-Value: <2.2e^-16

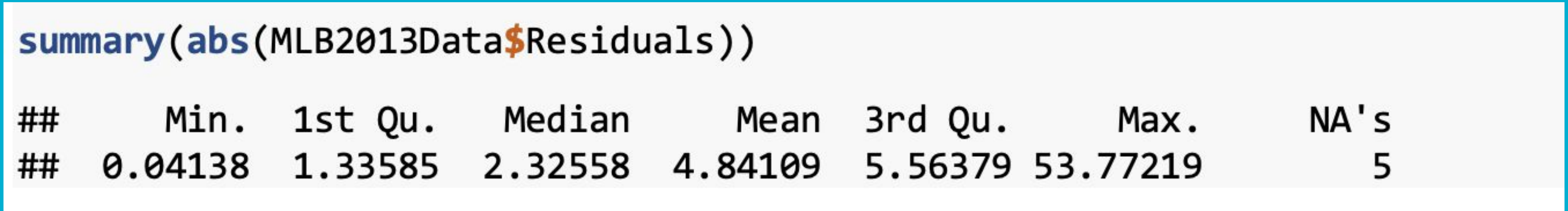
# Testing Our Model: 2013 MLB Season

- Using AIC_Model2, we predicted the number of wins each MLB team in 2013 season.
- The code we ran is shown below:



```
PredictedWins=predict(AIC_Model2,newdata=MLB2013Data)
MLB2013Data$PredictedWins=PredictedWins
MLB2013Data$Residuals=(MLB2013Data$W - MLB2013Data$PredictedWins)
print(MLB2013Data[,c("Team","W","PredictedWins","Residuals")])
```

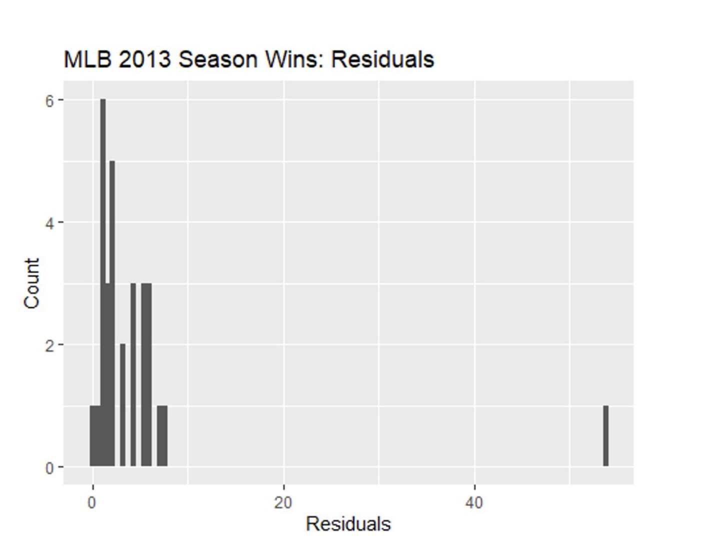
# Testing Our Model: 2013 MLB Season



```
summary(abs(MLB2013Data$Residuals))

##      Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##  0.04138  1.33585  2.32558  4.84109  5.56379 53.77219        5
```

AIC_Model2 predicted the number of wins with a mean residual value of 4.84109 and the median residual value was 2.32558.

# Testing Our Model: 2013 MLB Season



AIC_Model2 had a mean absolute value residual of 4.84109:

On average, our predicted number of games won for each team in the 2013 MLB season was off by an absolute value of 5.
The median residual was 2.32558 meaning that we predicted 50% of the teams total wins by a margin of plus or minus 2.32558.
In summary, our model predicted the total number of wins a team will have over the course of a season very well.