Final Project

Jordan Badstuebner

May 2020

Introduction

Baseball is a game dear to my heart. It is a game of many variables, many of which can be measured, recorded, and analyzed and which have been for over a century. For the purpose of this project, being the icing to my "curiosity as a fan of the game" cake, I seek to use this data, which contains some simple baseball statistics, to explore the possibility any of those basic statistics correlate to winning baseball games, as well as the possibility of producing a reasonable prediction model.

Summary of Data

The data set used for this experiment is easily accessible via the internet and titled "baseball.csv" A preview and summary of the data set are shown below:

As can be seen in **Table 1**, rows comprise all teams in Major League Baseball (MLB) and columns contain some common baseball stats. In **Table 2**, we can see that the data covers years 1962-2012. For the purposes of this analysis, we will focus on wins (W) as the dependent variable, and independent variables: runs scored (RS), runs agains (RA), onbase percentage (OB), slugging (SLG), batting average (BA), opposing OBP (OOBP), and opposing SLG (OSLG).

The Question

The question is simple - What are the factors that can be shown to correlate to winning?

Data Exploration

Before looking at any models, I ran the sum of (RS) against the sum of (RA). The sums of these columns were indeed equivalent, which is a good sign the data was input correctly into these columns.

Table 1: Preview of Baseball.csv

Team	League	Year	RS	RA	W	OBP	SLG	BA	Playoffs	RankSeason	RankPlayoffs	G	OOBP	OSLG
ARI	NL	2012	734	688	81	0.328	0.418	0.259	0	NA	NA	162	0.317	0.415
ATL	NL	2012	700	600	94	0.320	0.389	0.247	1	4	5	162	0.306	0.378
BAL	AL	2012	712	705	93	0.311	0.417	0.247	1	5	4	162	0.315	0.403
BOS	AL	2012	734	806	69	0.315	0.415	0.260	0	NA	NA	162	0.331	0.428
CHC	NL	2012	613	759	61	0.302	0.378	0.240	0	NA	NA	162	0.335	0.424
CHW	AL	2012	748	676	85	0.318	0.422	0.255	0	NA	NA	162	0.319	0.405

Table 2: Summary of Baseball.csv

Team	Year	RS	RA	W	OBP	SLG	BA	OOBP	OSLG
Length:1232	Min. :1962	Min. : 463.0	Min. : 472.0	Min. : 40.0	Min. :0.2770	Min. :0.3010	Min. :0.2140	Min. :0.2940	Min. :0.3460
Class :character	1st Qu.:1977	1st Qu.: 652.0	1st Qu.: 649.8	1st Qu.: 73.0	1st Qu.:0.3170	1st Qu.:0.3750	1st Qu.:0.2510	1st Qu.:0.3210	1st Qu.:0.4010
Mode :character	Median :1989	Median : 711.0	Median : 709.0	Median : 81.0	Median :0.3260	Median :0.3960	Median :0.2600	Median :0.3310	Median :0.4190
NA	Mean :1989	Mean : 715.1	Mean : 715.1	Mean : 80.9	Mean :0.3263	Mean :0.3973	Mean :0.2593	Mean :0.3323	Mean :0.4197
NA	3rd Qu.:2002	3rd Qu.: 775.0	3rd Qu.: 774.2	3rd Qu.: 89.0	3rd Qu.:0.3370	3rd Qu.:0.4210	3rd Qu.:0.2680	3rd Qu.:0.3430	3rd Qu.:0.4380
NA	Max. :2012	Max. :1009.0	Max. :1103.0	Max. :116.0	Max. :0.3730	Max. :0.4910	Max. :0.2940	Max. :0.3840	Max. :0.4990
NA	NA	NA	NA	NA	NA	NA	NA	NA's :812	NA's :812

The second step was was to check for outliers. I started with (W). The box plot titled **Outliers of Response Variable** is shown below and **Table 3** provides a great slice of the teams and numbers corresponding to those teams' respective outlier season. There exist only 4 outlies out of 880,981 observations. The histogram below titled **Distribiution of Response Variable** shows a normal distribution around the mean of 80.9.

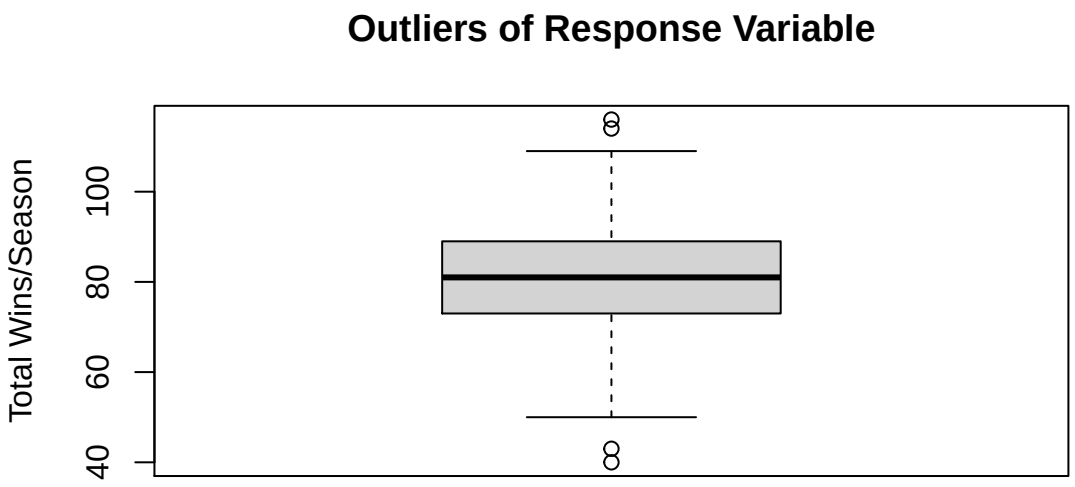
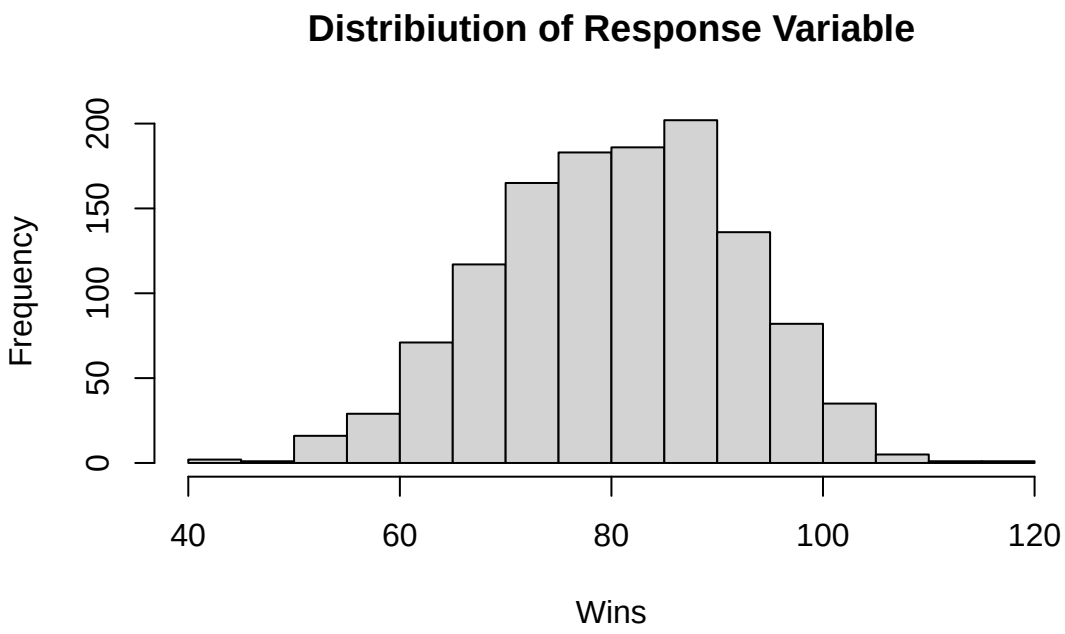


Table 3: (W) Outliers slice

	Team	League	Year	RS	RA	W	OBP	SLG	BA	Playoffs	RankSeason	RankPlayoffs	G	OOBP	OSLG
281	DET	AL	2003	591	928	43	0.300	0.375	0.240	0	NA	NA	162	0.352	0.461
355	SEA	AL	2001	927	627	116	0.360	0.445	0.288	1	1	3	162	0.301	0.378
440	NYY	AL	1998	965	656	114	0.364	0.460	0.288	1	1	1	162	NA	NA
1226	NYM	NL	1962	617	948	40	0.318	0.361	0.240	0	NA	NA	161	NA	NA



The Model

As a position player (not a pitcher), my initial thoughts went back to what any postion player has heard from his coach since little league - "hits hits hits!" My initial guess was that batting average will play a huge roll in winning. That linear model returned a p-value of 2e-16, showing that batting average is definitely significant, but the R^2 of this model was an abismal 0.1664.

After giving this some follow up thought, I realized (OBP) is a smarter replacement for (BA) and more likely to possess a stronger correlation to winning, given that (BA) is factored into (OBP). Running another linear model with (OBP) as a the independent variable, the return was, again, a significant p-value with a very low R^2 and, of course, this makes sense. Getting on base is one thing, but runs must be scored to win.

Having now found that (OBP) possesses a similar p-val and higher R^2 than (BA), the next step was to use a multiple regression and factor (OBP) with (SLG). Surely, (OBP) and (SLG) must correlate to winning. Couriously, (SLG) returned a p-val of 0.182, showing it wasn't significant to this model at all. So, (OBP) is signficant to winning and has a [wicked] weak correlation to winning, but why? Well, this actually makes a lot of sense.

What is the reward for high (OBP) and (SLG)? The reward is (RS). Sure enough, after using (RS) as the depencant variable and (OBP)+(SLG) as independant variables, we are now cooking with fire. Both (OBP) and (SLG) returned p-values of 2e-16 and the model returned an R^2 of 0.9247.

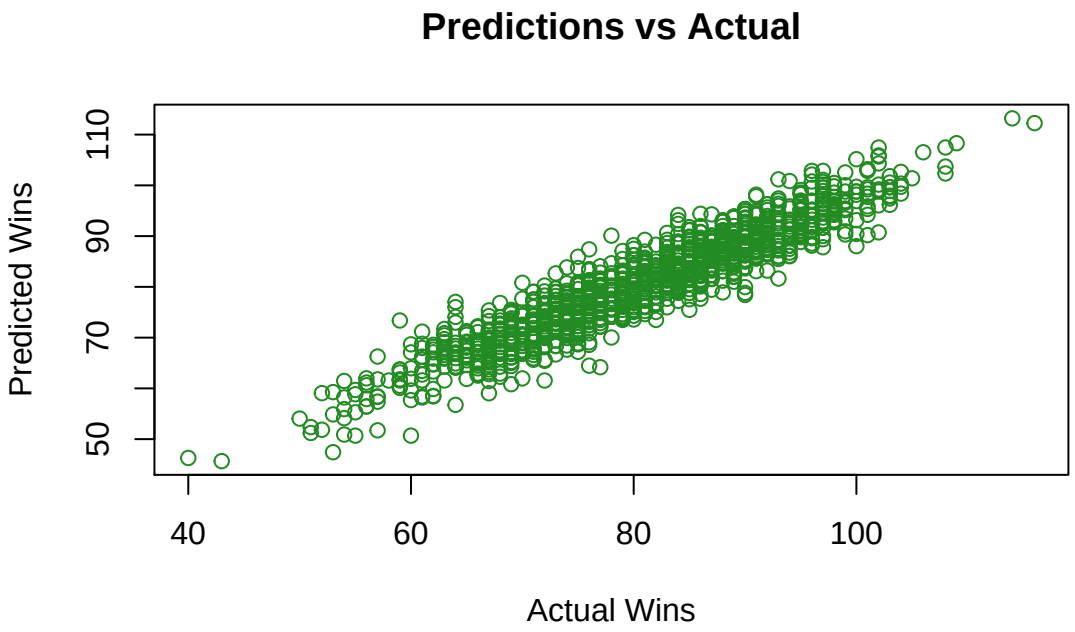
Returning to (W) as the response variable, I ran the model with (RS) which returned a highly significant p-value of 2e-16 but another heineous R^2 of .2613. So (OBP)+(SLG) is a strong correlation to (RS), which is highly signicant to (W), but has a weak correlation to (W). And then it *hit* me:

```
##
## Call:
## lm(formula = df$W ~ df$RS + df$RA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3776  -2.7753   0.0513   2.8051  12.8298
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 80.980456   1.063978   76.11   <2e-16 ***
## df$RS        0.104493   0.001340   78.00   <2e-16 ***
## df$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
```

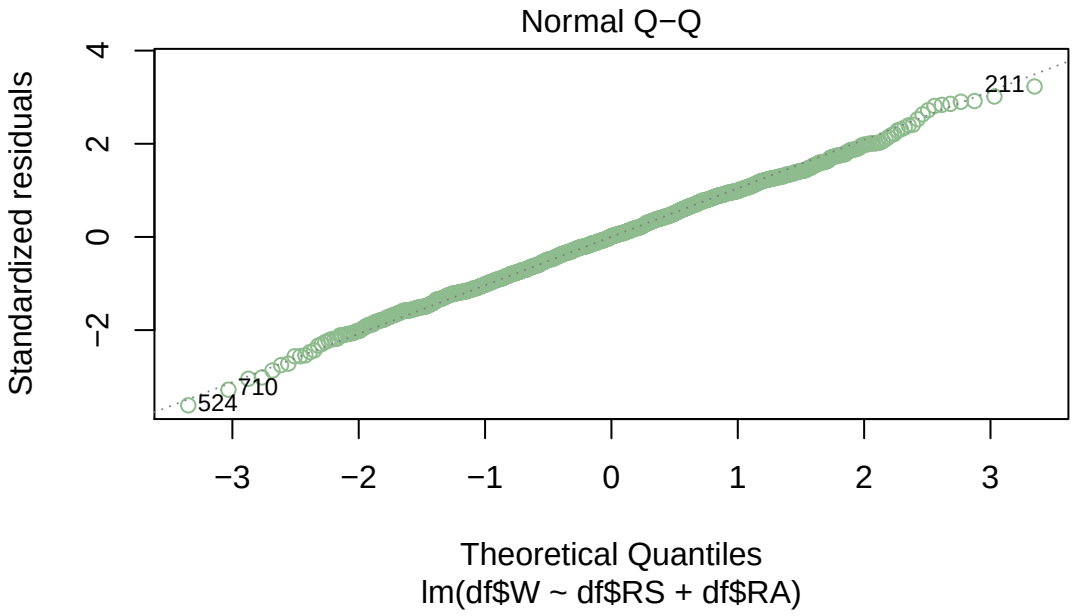
How does one win a baseball game, or any scored match for that matter? Score a lot of points and don't give up a lot of points, of course. It appears that defense and pitching actually matter too! As shown above, R^2 returns at 0.8794, giving the model a respectible proportion of variability explained by the regression (it's a decent model). (RS) and (RA) show highly significant p-values of 2e-16 at >99.999% confidence. (RS) returns a slope of 0.104493 telling us that for every run scored, wins increase by ~.104. (RA) returns a negative slope (gratefully) of -0.104600 telling us that for every run scored against a team, wins decrease by ~.104. I found it interesting, and worth noting, that these slopes are nearly identical.

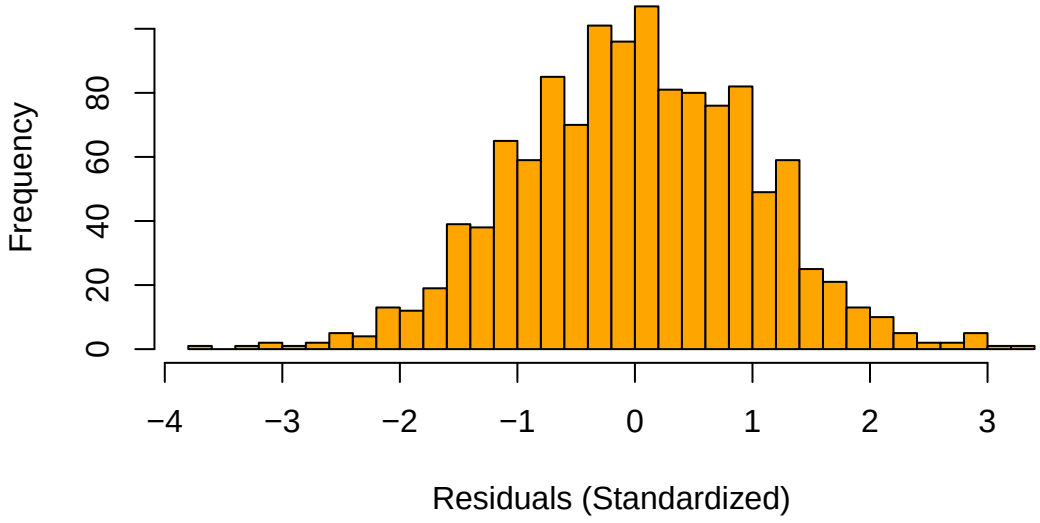
Residual Analysis

As indicated in the plot below, the predictions of the model conform to the diagonal normality line - a good sign we are making valid inferences from our regression.

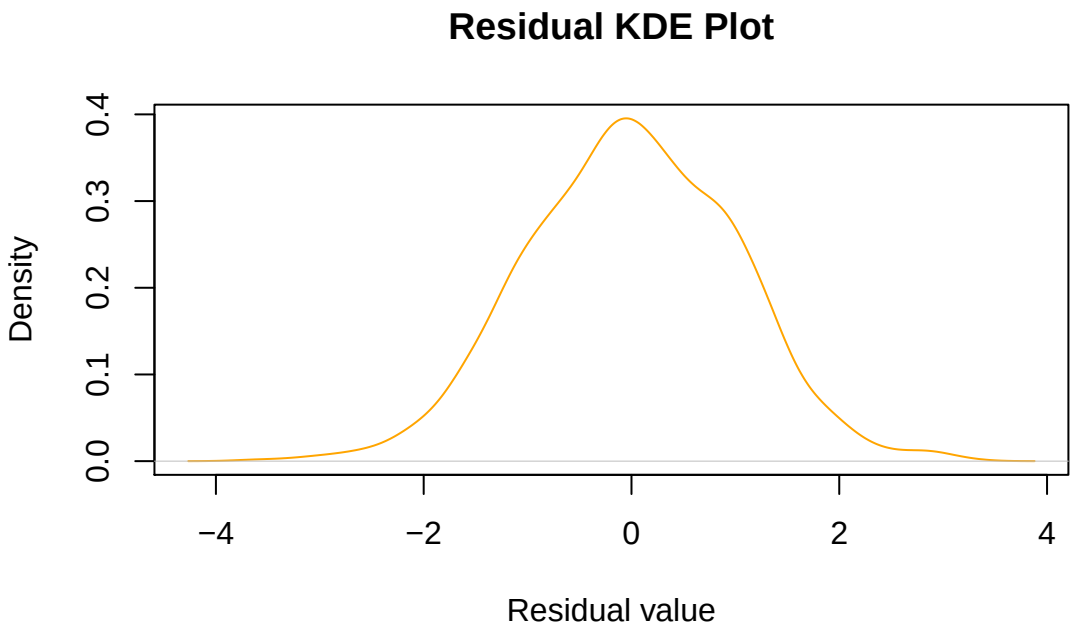


I proceeded to standarize the residuals to give us a better understanding of them. Taking a closer look into the normal q-q, histogram, and desity plots below, our residuals appear to be normally distributed.

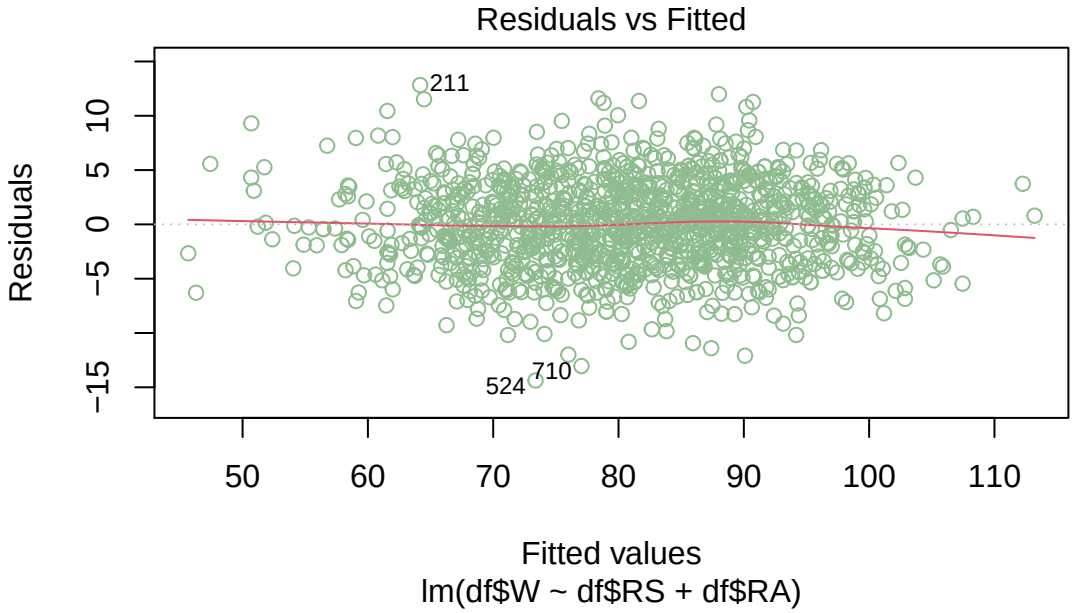


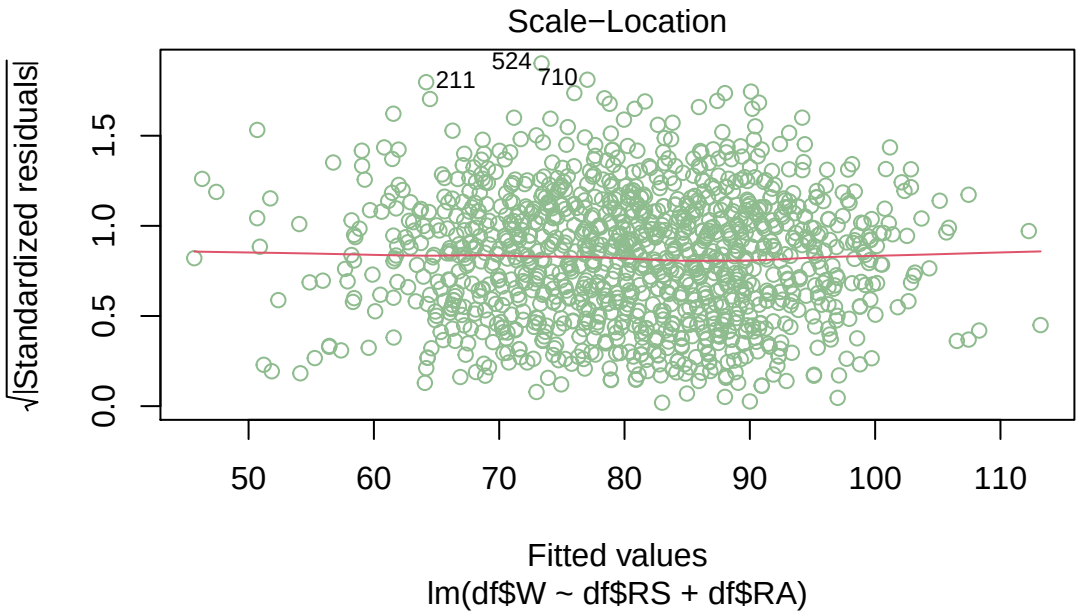



```
## $breaks
##  [1] -3.8 -3.6 -3.4 -3.2 -3.0 -2.8 -2.6 -2.4 -2.2 -2.0 -1.8 -1.6 -1.4 -1.2 -1.0
## [16] -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0  1.2  1.4  1.6  1.8  2.0
## [31]  2.2  2.4  2.6  2.8  3.0  3.2  3.4
##
## $counts
##  [1]   1   0   1   2   1   2   5   4  13  12  19  39  38  65  59  85  70 101  96
## [20] 107  81  80  76  82  49  59  25  21  13  10   5   2   2   5   1   1
##
## $density
##  [1] 0.004058442 0.000000000 0.004058442 0.008116883 0.004058442 0.008116883
##  [7] 0.020292208 0.016233766 0.052759740 0.048701299 0.077110390 0.158279221
## [13] 0.154220779 0.263798701 0.239448052 0.344967532 0.284090909 0.409902597
## [19] 0.389610390 0.434253247 0.328733766 0.324675325 0.308441558 0.332792208
## [25] 0.198863636 0.239448052 0.101461039 0.085227273 0.052759740 0.040584416
## [31] 0.020292208 0.008116883 0.008116883 0.020292208 0.004058442 0.004058442
##
## $mids
##  [1] -3.7 -3.5 -3.3 -3.1 -2.9 -2.7 -2.5 -2.3 -2.1 -1.9 -1.7 -1.5 -1.3 -1.1 -0.9
## [16] -0.7 -0.5 -0.3 -0.1  0.1  0.3  0.5  0.7  0.9  1.1  1.3  1.5  1.7  1.9  2.1
## [31]  2.3  2.5  2.7  2.9  3.1  3.3
##
## $xname
## [1] "rstandard(f)"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```



In order to evalue the homoscedasticity of the residuals, I used the graphs below. In both graphs, it is reasonable to determine a completely random equal distribution of points throughout the range of x along what is very close to a flat red line. This indicates there is almost no heteroscedastity.





Since our residuals are nomrally distributed and homoscedastic, we need not evaluate linearity. Concluding our residual analyis, we can be confident that we are making valid inferences from our regression.

Conclusion

It may seem intuitive that the key to winning is to maintain the highest possible number of runs scored in tandem with allowing the lowest possible number of runs to be scored against you. With this model, we are able to go beyond intuition, state that (RS) and (RA) is strongly correlated to winning in baseball, and even predict wins based on a team's (RS) and (RA) ratio at any point in a season. After analyzing the data in deeper depth, there were potentially strong correlations of:

(OBP) and (SLG) to (RS) | p-val<.001, R^2>.9

(OOBP) and (SLG) to (RA) | p-val<.001, R^2>.9

These results are, again, intuitive and it would be safe to bet that they can be determined as correlations following a repeat process of this analysis.

There are a mindblowing amount of statistics recorded in the game of baseball. As techology continues to improve, so will the the invention of new and exciting sports statistics. This data set was limited in that it only included some of the most basic statistics in the game. I would have loved to have taken a look at things like fielding percentage, atompheric pressure and tempurature, or designated list times as a few examples of additional factors. I am curious to see what factors might show a stronger correlation once added into our model.

I am now a coach. This summer, while stressing to the hitters to get on base, then stressing to the pitchers to throw strikes, it will not be because "I say so", but, keeping the findings of this lab in mind, it will now be because I have a pretty good idea that they are part of a strong correlation.