Analysis of the 2003-2018 Major League Baseball Regular Seasons

STAT 510 Final Report

Mario A. Leon
DoKyoung Shin

California State University, Long Beach
December 17, 2021

Background

This data set shows the regular season statistics for each Major League Baseball team from 2003-2018. Postseason statistics are not considered. During the 2003-2018 period, the Boston Red Sox won the most World Series Championships with 4 (2004, 2007, 2013, 2018), followed by the San Francisco Giants with 3 (2010, 2012, 2014) and the St. Louis Cardinals with 2 (2006, 2011). In regards to the every World Series champion from 2003-2018, only two champions out of the 16 had less than 90 wins in the regular season so a good threshold for a successful team is normally about 90 wins. Having 90 wins gives a team a good chance to qualify for the postseason and compete for the World Series title.

With that thought, we were interested in conveying which factors contribute to a team's success in the regular season, using W (Total Team Wins) as our response variable. Some of the variables we have include WAR (Wins Above Replacement, which determines how valuable a player is), salary, ERA (Earned Run Average), SO (Strikeouts Against), SO 1 (Strikeouts For), BA (Batting Average), H (Hits Allowed), RBI (Runs Batted In), H1 (Hits For), E (Errors) and many others. This may be able to give us an idea as to why the Red Sox, Giants, and Cardinals were dominant during this time period.

Questions of Interest

In order to delve further into our primary focus, it is first important to ponder about how basic statistics such as hitting and pitching affect a team's season. It is often discussed by baseball enthusiasts that good hitting and good pitching equals a good team so we were interested in how many certain batting average (BA) and earned-run average (ERA) affect a team's win total without taking other variables into account.

When we build our model, we will be interested in what other variables besides good hitting and good pitching affect a team's win total. Once we have obtained those variables in our final model, we would like to see how much of the variation in Wins is explained by the predictors we obtain. Continuing with the final model, we are also interested in analyzing whether a team's win total is significantly related to both the offensive and defensive statistics after controlling for all other predictors in our final model. We also want to determine whether salary is significant to a team's Win total since it is often discussed that a team with a lot of money can "purchase wins" by signing the best players available.

Major League Baseball currently has 30 teams divided into two leagues called the American League (AL) and National League (NL), with 15 teams in each league. The World Series match-up is also between the AL champions vs the NL champions, but it may be possible that one team had a much easier path than their opposition. A possible question of interest will be how much different the average win total of an NL team versus an AL team is. Unfortunately, there is no variable to represent what league each team plays for. We will take care of this problem later.

Analysis

1 How many wins can expect on an average team given specific ERA and BA results?

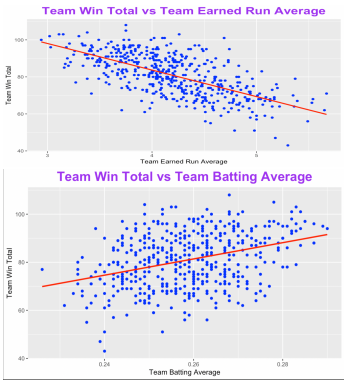
We will only work with the model $Wins=\beta_0+\beta_1 ERA+\beta_2 BA$ in this case, which we call the "Casual" model because any typical baseball enthusiast constitutes a successful baseball team to be a team that can hit and pitch without taking any other factors into consideration. In this case, the only variables worth looking at are batting average (BA) and earned-run average (ERA) for hitting and pitching, respectively. This model also satisfies normality, linearity, and constant variance (Appendix Fig. 3).

Running a 95% confidence interval for a team with a 5.00 ERA and a .250 BA, we are 95% confident that this team will have between 61 and 64 wins, which is not good. This makes sense since the higher the ERA, the worse a team will perform as the relationship between ERA and W is negative. A successful team typically has their ERA as low as possible as that shows they do not allow runs to score as often.

If we happened to improve ERA to 2.50 but lower BA by .175, we are now 95% confident that this team will have around 63 and 72 wins, which is better than before but not by much as this is still not a playoff team. The ERA may be better but if the team cannot hit, then the chances of winning are still not good.

Improving ERA to 2.50 and BA to .250 makes us 95% confident that this team will have 103 to 107 wins, meaning we expect this team to be in World Series contention. We see that good hitting and good hitting have an effect on a team's success but now we want to look at what other factors contribute to a team's win total.

Fig. 1. Wins vs. Earned-Run Average and Wins vs. Batting Average



Model-Building Process

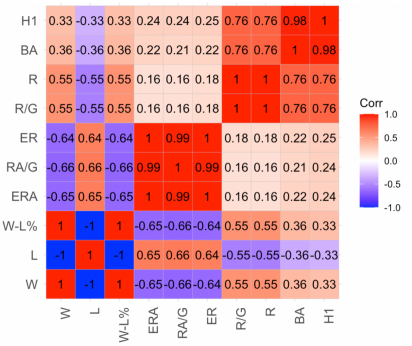
The variables given in the data set have the Year and the Team combined into one variable called YearTeam. However, we would like to know if the Year and Team separately have anything to do with a successful team so we will split them up into two separate variables (Code 2). This will also facilitate our process of creating the League variable as all we need to do is assign the teams that played in the American League (AL) and National League (NL) during this time period.

The AL teams are represented with a "0" and NL with a "1". A couple of caveats here are the Angels being the "Anaheim Angels" from 2003-2004 ("ANA") and the "Los Angeles Angels" from 2005-2018 ("LAA"). The same idea is applied for Tampa Bay, who were the "Tampa Bay Devil Rays" from 2003-2007 ("TBD") and the "Tampa Bay Rays" ("TBR"). Finally, we need to account for the Houston Astros being part of the NL from 2003-2012 and the AL from 2013-2018. We now

have created our new "league" variable (Code 3) and a data set with "Year", "Team", and "League" (Appendix Fig. 1).

Now that we have our data set in good condition, there are other caveats to consider about the predictors we have. It turns out that some variables are already giving us the same information. If we do not address the multicollinearity issue, we will get a very futile model. The data set contains ERA, ER, and RA/G, which all tell us the runs allowed by a team but in different measurements. We will use ERA as it is the most prominent statistic used by sabermetricians. Other redundant cases include having W, L, and W-L% (keep W) and R and R/G (keep R) as shown by their perfect correlation. However, we will keep H1 and BA despite the high correlation because this may tell us different information about how efficient a team's offense is. Only tSho (total Shutouts) will be excluded out of our own judgment as we are not interested in what the score was, whether it was 2-1 or 2-0. We only care that the end result was a win. Hence, the excluded variables will be RA/G, ER, R/G, L, W-L%, and tSho. This should allow us to have a more practical model with several unique factors to consider.

Fig. 2. Correlation between the variables we considered excluding.



We decided to use step-wise AIC for variable selection since it produced a more practical model compared to the other variable selection techniques. The step-wise selection selected variables until it reached an AIC value of 1314.99. The predictors chosen by AIC are WAR, League, Runs, ERA, E, OBP, salary, Hits (H1), BA, and Hits Allowed (H) (Code 4), with W as our response variable. Based on this model, we can see its characteristics, such as: residual standard error is 3.89, multiple R-squared is 0.8851 and adjusted R-squared is 0.8826, with all the predictors significant **(Fig.3)**.

Fig. 3. Summary of our AIC Model

```
Call:
lm(formula = W ~ WAR + League + R + ERA + E + OBP + salary +
    H1 + BA + H, data = MLBTeamStats1)

Residuals:
     Min       1Q   Median       3Q      Max
-11.0956  -2.6740  -0.1651   2.4333  12.4369

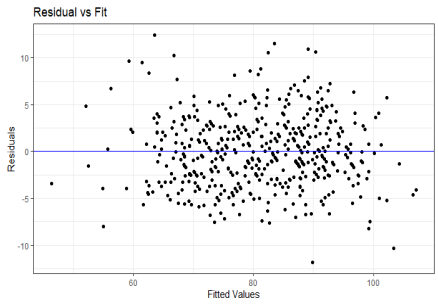
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.153e+01  7.470e+00   6.898 1.72e-11 ***
WAR          3.277e-01  7.152e-02   4.581 5.93e-06 ***
League1      1.376e+00  5.101e-01   2.698 0.007218 **
R            7.434e-02  7.152e-03  10.394  < 2e-16 ***
ERA         -1.223e+01  1.173e+00 -10.427  < 2e-16 ***
E           -4.615e-02  1.316e-02  -3.507 0.000496 ***
OBP         -7.131e+01  3.488e+01  -2.044 0.041486 *
salary       1.060e-08  4.325e-09   2.451 0.014606 *
H1          -8.188e-02  1.351e-02  -6.061 2.79e-09 ***
BA           5.576e+02  9.545e+01   5.841 9.70e-09 ***
H            1.017e-02  3.923e-03   2.592 0.009842 **
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.89 on 469 degrees of freedom
Multiple R-squared:  0.8851,    Adjusted R-squared:  0.8826
F-statistic: 361.3 on 10 and 469 DF,  p-value: < 2.2e-16
```

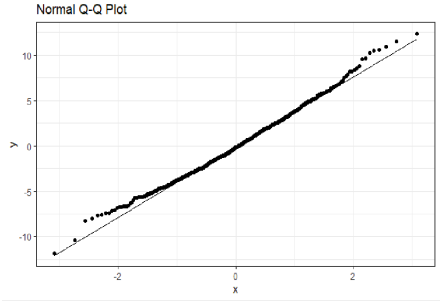
We will now check for potential significant interaction terms involving salary. For some background on why salary was chosen, it is common to see teams paying high prices to acquire the best players they could possibly afford in hopes of making that team successful. In our data, the LA Dodgers were the richest team from 2013 to 2017. Based off real life circumstances, it is worth checking how much of an interaction salary has on certain variables. This may demonstrate what most of the money is going to. Using the *add1* function, we found that the significant interaction terms are *R*salary* and *OBP*salary* (Code 6).

When we add these two significant interaction terms (Code 7), it turns out that OBP and the interaction between R and salary are no longer significant, so we will remove these two variables

from our model (Code 8). Next, we will check the diagnostics of this model.



(a) Residual vs fitted plot



(b) Normal Q-Q plot

Fig. 4. The diagnostics of $Wins=\beta_0+\beta_1 WAR+\beta_2 League+\beta_3 ERA+\beta_4 E+\beta_5 H1+\beta_6 BA+\beta_7 H+\beta_8 R+\beta_9 salary+\beta_{10} OBP*salary$.

The Residuals vs Fitted plot of our new model **(Fig.4-(a))** demonstrates a well-behaved pattern as the points bounce randomly around the zero. This concludes that there is the linear relationship. In addition, the residuals roughly form a horizontal band around zero, meeting the constant variance assumption. Consequently, the residual versus fitted plot shows non-constant variance and non-linearity issue in our model.

The testing that we are using assume the null hypothesis states that the error terms are normally distributed, against the alternative hypothesis that they are not normally distributed. To check this condition, we used the normal quantile-quantile (Q-Q) plot **(Fig.5-(b))**. Notice the points seem to fall along a straight line, so it appears to be a fairly satisfy our assumption. We can also use the Shapiro-Wilk normality test to check normality, which gave us a p-value=0.03965. We had initially attempted to transform this several times to see if we can get a Shapiro p-value above 0.05 but after many unsuccessful attempts, we determined that transformations were superfluous and settled on concluding normality at a 0.01 level of significance. This is about what we expected as the distribution of Wins is approximately normal, aside from a couple of poor-performing teams (Appendix Fig. 2).

Fig. 5. Summary of Final Model

```
Call:
lm(formula = W ~ WAR + League + ERA + E + H1 + BA + H + R + salary +
    OBP:salary, data = MLBTeamStats1)

Residuals:
     Min       1Q   Median       3Q      Max
-11.8249  -2.7397  -0.1435   2.4860  12.4066

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.755e+01  8.638e+00   3.189 0.001522 **
WAR          3.446e-01  7.082e-02   4.866 1.56e-06 ***
League1      1.421e+00  4.928e-01   2.884 0.004110 **
ERA         -1.193e+01  1.170e+00 -10.198  < 2e-16 ***
E           -4.203e-02  1.314e-02  -3.199 0.001474 **
H1          -8.346e-02  1.323e-02  -6.310 6.48e-10 ***
BA           5.671e+02  9.097e+01   6.234 1.01e-09 ***
H            9.210e-03  3.858e-03   2.387 0.017360 *
R            7.375e-02  6.787e-03  10.866  < 2e-16 ***
salary       2.374e-07  6.756e-08   3.515 0.000483 ***
salary:OBP  -6.934e-07  2.065e-07  -3.358 0.000849 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.861 on 469 degrees of freedom
Multiple R-squared:  0.8868,    Adjusted R-squared:  0.8844
F-statistic: 367.4 on 10 and 469 DF,  p-value: < 2.2e-16
```

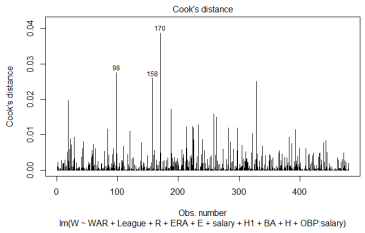
Therefore, our final model is:

Wins=β_0+β_1WAR+β_2League+β_3ERA+β_4E+β_5H1+β_6BA+β_7H+β_8R+$\beta_9$salary+β_{10}OBP*salary

We have decided to use Cook's distance to identify influential points. If the Cook's distance measure of an observation is greater than 0.5, then that observation could potentially be influential. According to our Cook's distance plot **(Fig.6)**, observations 170, 98, and 158 are

points to consider. However, since those points have a distance less than 0.5, so we can assume that those points are not influential. Even when we compared the two regression summaries (Appendix Fig.5), removing these three points from the model has slight change from our original model. The sign and magnitude of the slope parameter estimates are very similar. The adjusted R-squared of the removed observations model is 0.8906 that is slightly better than our model but all of these changes are not too drastic. Thus, we can conclude that these three points are not influential. We are content with having the model with these three observations.

Fig. 6. Cook's Distance



2 How much variation in Wins is explained by our Final Model?

According to the R^2 in our final model, 88.68% of the variation in Wins is explained by WAR, League, ERA, Errors, Hits (H1), Hits Allowed (H), Batting Average, salary, Runs, and the interaction between salary and On-Base Percentage, which is superb.

3 Is a team's Win total significantly related to the offensive/defensive statistics after controlling for all other predictors in our model?

In the data set, the offensive statistics are BA, H1, R, and interaction between OBP and salary. To answer this research question, we test the hypotheses:

H_0: $\beta_5 = \beta_6 = \beta_8 = \beta_{10} = 0$

H_a: At least one $\beta_i \neq 0 (i = 5, 6, 8, 10)$

The full model contains all possible predictors from our model while the reduced model excludes BA, H1, R and OBP*salary. From the analysis of variance table (ANOVA) **(Fig.7)**, F-statistic value is 36.99, which is greater than critical value 3.86 and the p-value is very small. Therefore, we reject the null hypotheses and conclude that the offensive statistics from our model are significantly related to how successful a team is after controlling for everything else.

Fig. 7. ANOVA table for offensive (above) and defensive (below) statistics

```
Analysis of Variance Table

Model 1: W ~ WAR + League + ERA + E + H + salary
Model 2: W ~ WAR + League + ERA + E + H1 + BA + H +
R + salary + OBP:salary
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    473 9198.6
2    469 6992.8  4    2205.9 36.987 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of Variance Table

Model 1: W ~ WAR + League + H1 + BA + R + salary +
OBP:salary
Model 2: W ~ WAR + League + ERA + E + H1 + BA + H +
R + salary + OBP:salary
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    472 8566.2
2    469 6992.8  3    1573.5 35.177 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, we will perform the same process but with the defensive statistics ERA, Errors, and H. The hypotheses are:

H_0: $\beta_3 = \beta_4 = \beta_7 = 0$

H_a: At least one $\beta_i \neq 0 (i = 3, 4, 7)$

The full model contains all possible predictors from our model while the reduced model excludes ERA, E, and H. From ANOVA table **(Fig.7)**, statistic F value is 35.18, which is greater than critical value of 3.86 and the p-value is very small. Therefore, we can

reject the null hypotheses and conclude that the defensive statistics from our model are significantly related to how successful a team is after controlling for everything else. Both offense and defense are confirmed to be important! Not one or the other.

4 **Is Salary also significant? Does salary have a significant interaction term?**

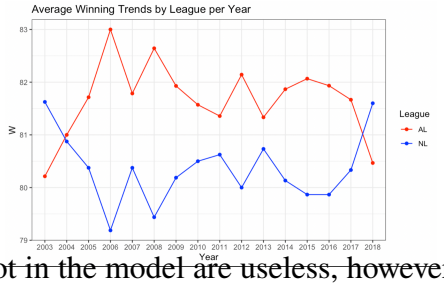
Based off our summary output, we see that Salary and the interaction between Salary and On-Base Percentage are both significant terms since the p-values for both terms are extremely close to zero. Hence, we can say that salary plays an important role in determining how many wins a team will get as the money could be used to sign top players to help lead the team to success. The interaction being significant also makes sense as the effect of on-base percentage on Wins is influenced by salary. In other words, we can say that money seems to be going towards players who can get on base. This same philosophy was referenced in the 2011 film "Moneyball".

5 **How much different is the average Win total of an NL team vs an AL team?**

Fortunately, the League variable we created earlier ended up being significantly part of our model so we can answer this question. When we run a 95% confidence interval on the coefficients of our model, the result for the League variable tells us that we are 95% confident that the average regular season wins of NL teams is between 0.45 and 2.839 more than the average regular season wins of AL teams regardless of the other predictors. When we graph the average trends of teams in each league per season, we see an increase in NL teams and a decrease in AL wins overtime as shown in **Fig. 8**. To ponder why that may be the case, perhaps the competition in the NL got better or the NL teams are emphasizing something that AL teams are not.

Conclusion

Fig. 8. The average wins of AL teams are declining overtime as the NL is increasing, which is accurate with our Confidence Interval interpretation.



Based off our analysis, we have determined that Wins Above Replacement, League, ERA, Errors, Hits, Hits Allowed, Batting Average, salary, Runs, and the interaction between Salary and On-Base Percentage are significant for a successful MLB team. To potentially enhance this model, we could have created other variables that were not present in the data set in addition to League, such as WHIP (Walks-Hits per Innings Pitched) and OPS (On-Base plus Slugging Percentage). Right now, we can definitely state that a team might be successful because the league they play in might fit their strengths better. This does not mean that the variables not in the model are useless, however. RBI and LOB, for example, could possibly be a concern for a team that they need ameliorate. Every team has their own unique strengths and weaknesses but we can conclude that it is salient to consider where a team's weakness is amongst the statistics shown in our final model and improve on that.

Appendix

R Code

(Code 1) Analyzing only hitting and pitching: the values of ERA and BA simply changed to see what would happen.

```
mod1 <- lm(W ~ERA + BA, data = MLBTeamStats)
new1 <- data.frame(ERA = 5.00, BA = .250)
predict(mod1, new1, interval = "confidence")
```

(Code 2) Splitting up YearTeam into Year and Team

```
MLBTeamStats1 <- MLBTeamStats %>% separate(YearTeam, c("Year", "Team"), " ")
```

(Code 3) Creating the variable League

```
MLBTeamStats1$League = ifelse(MLBTeamStats1$Team %in% c("LAA", "TEX", "OAK", "SEA", "MIN", "KCR", "DET", "CHW", "CLE", "BOS", "NYY", "TBR", "TOR", "BAL", "ANA", "TBD"),0, ifelse(MLBTeamStats1$Team %in% c("HOU") & MLBTeamStats1$Year >= 2013, 0, 1))
MLBTeamStats1$League <- as.factor(MLBTeamStats1$League)
```

(Code 4) Variable selection - Step-wise AIC model

```
mod0 = lm(W ~1, data = MLBTeamStats1)
mod.all = lm(W ~. -'RA/G' -'L -W-L%' -tSho -ER -'R/G', data = MLBTeamStats1)
step(mod0, scope = list(lower= mod0, upper = mod.all))
```

(Code 5) Summary of AIC model

```
mod.AIC = lm(W ~WAR + League + R + ERA + E + OBP + salary + H1 + BA + H, data = MLBTeamStats1)
summary(mod.AIC)
```

(Code 6) Interaction term with Salary

```
add1(mod.AIC, ~.  + salary*WAR + salary*League + salary*R + salary*ERA + salary*E + salary*OBP + salary*H1 + salary*BA + salary*H, test = 'F')
```

(Code 7) Interaction term with Salary - Adding two interaction terms

```
mod.2 = update(mod.AIC, ~. + salary*R+ salary*OBP)
summary(mod.2)
```

(Code 8) Interaction term with Salary - Removing insignificant variables

```
mod.3 = update(mod.2, ~. - OBP - salary*R + salary + R)
summary(mod.3)
```

(Code 9) Creating the table for the model

```
mod_table3 = augment(mod.3)
```

(Code 10) The diagnostics of the final model - Residuals vs Fitted plot

```
ggplot(mod_table3, aes(x = .fitted, y = .resid)) +
geom_point() + geom_hline(yintercept = 0, colour = 'blue') +
labs(x = 'Fitted Values', y = 'Residuals') + ggtitle('Residual vs Fit') + theme_bw()
```

(Code 11) The diagnostics of the final model - Normal Q-Q plot and Shapiro-Wilk test

```
ggplot(mod_table3, aes(sample = .resid)) + stat_qq() + stat_qq_line() +
ggtitle('Normal Q-Q Plot') + theme_bw()
shapiro.test(resid(mod.3))
```

(Code 12) Cook's distance plot

```
plot(mod.3,which =4)
```

(Code 13) Summary plot of after deleting possible influential points

```
mod.3_delete = lm(W ~WAR + League + R + ERA + E + salary + H1 + BA + H + OBP*salary,
data = MLBTeamStats1[-c(98, 158, 170), ])
```

(Code 14) Relationship between a team's Win total and offensive statistics

```
mod_reduced1 = lm(W ~WAR + League + ERA + E + H + salary, data = MLBTeamStats1)
mod_full1 = lm(W ~WAR + League + ERA + E + H1 + BA + H + R + salary + OBP*salary,
data = MLBTeamStats1)
anova(mod_reduced1, mod_full1)
```

(Code 15) Relationship between a team's Win total and defensive statistics

```
mod_reduced2 = lm(W ~WAR + League + H1 + BA + R + salary + OBP*salary, data =
MLBTeamStats1)
mod_full2 = lm(W ~WAR + League + ERA + E + H1 + BA + H + R + salary + OBP*salary,
data = MLBTeamStats1)
anova(mod_reduced2, mod_full2)
```

(Code 16) 95% Confidence Interval of Coefficients of our Final model

```
confint(mod.3, level = 0.95)
```

Additional Plots

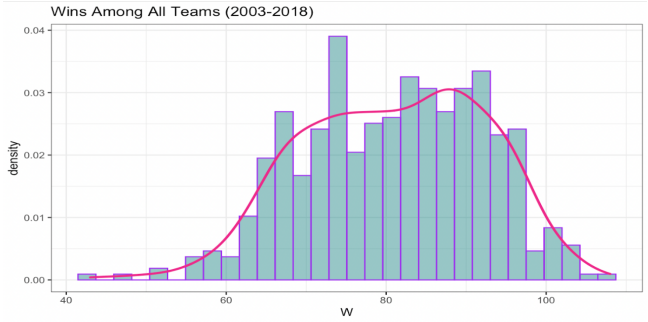
(Appendix Fig. 1) Variables before the changes (a) and variables after the changes (b).

```
```{r}
glimpse(MLBTeamStats)
```

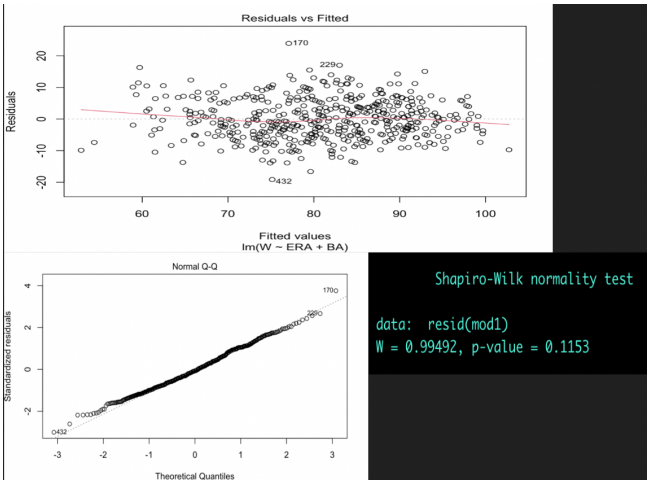
Rows: 480
Columns: 28
$ YearTeam <chr> "2009 ARI", "2009 ATL", "2009 BAL", "2009 BOS", "2009 CHC", "2009 CHW", "2009 CIN", "2…
$ `RA/G`   <dbl> 4.83, 3.96, 5.41, 4.54, 4.17, 4.52, 4.46, 5.34, 4.41, 4.57, 4.73, 4.75, 5.20, 4.70, 3.…
$ DefEff   <dbl> 0.683, 0.686, 0.680, 0.678, 0.698, 0.689, 0.704, 0.681, 0.688, 0.694, 0.686, 0.677, 0.…
$ E   <dbl> 124, 96, 90, 82, 105, 113, 89, 97, 87, 88, 106, 78, 117, 85, 83, 98, 76, 97, 86, 105, …
$ DP   <dbl> 133, 159, 151, 121, 144, 158, 161, 170, 146, 164, 129, 161, 159, 174, 134, 149, 135, 1…
$ W   <dbl> 70, 86, 64, 95, 83, 79, 78, 65, 92, 86, 87, 74, 65, 97, 95, 80, 87, 70, 103, 75, 93, 6…
$ L   <dbl> 92, 76, 98, 67, 78, 83, 84, 97, 70, 77, 75, 88, 97, 65, 67, 82, 76, 92, 59, 87, 69, 99…
$ `W-L%`   <dbl> 0.432, 0.531, 0.395, 0.586, 0.516, 0.488, 0.481, 0.401, 0.568, 0.528, 0.537, 0.457, 0.…
$ ERA   <dbl> 4.42, 3.57, 5.15, 4.35, 3.84, 4.14, 4.18, 5.06, 4.22, 4.29, 4.29, 4.54, 4.83, 4.45, 3.…
$ tSho   <dbl> 12, 10, 3, 11, 8, 11, 12, 6, 7, 9, 5, 10, 9, 13, 9, 8, 7, 12, 8, 10, 9, 7, 9, 10, 18, …
$ H   <dbl> 1470, 1399, 1633, 1494, 1329, 1438, 1420, 1570, 1427, 1449, 1425, 1521, 1486, 1513, 12…
$ ER   <dbl> 711, 581, 817, 695, 616, 663, 677, 806, 675, 690, 690, 722, 765, 715, 558, 770, 726, 7…
$ HR   <dbl> 168, 119, 218, 167, 160, 169, 188, 183, 141, 182, 160, 176, 166, 180, 127, 207, 185, 1…
$ BB   <dbl> 525, 530, 546, 530, 586, 507, 577, 598, 528, 594, 601, 546, 600, 523, 584, 607, 466, 6…
$ SO   <dbl> 1158, 1232, 933, 1230, 1272, 1119, 1069, 986, 1154, 1102, 1248, 1144, 1153, 1062, 1272…
$ `R/G`   <dbl> 4.44, 4.54, 4.57, 5.38, 4.39, 4.47, 4.15, 4.77, 4.96, 4.56, 4.77, 3.97, 4.23, 5.45, 4.…
$ R   <dbl> 720, 735, 741, 872, 707, 724, 673, 773, 804, 743, 772, 643, 686, 883, 780, 785, 817, 6…
$ H1   <dbl> 1408, 1459, 1508, 1495, 1398, 1410, 1349, 1468, 1408, 1443, 1493, 1415, 1432, 1604, 15…
$ RBI   <dbl> 686, 700, 708, 822, 678, 695, 637, 730, 760, 718, 727, 616, 657, 841, 739, 757, 770, 6…
$ SB   <dbl> 102, 58, 76, 126, 56, 113, 96, 84, 106, 72, 75, 113, 88, 148, 116, 68, 85, 122, 111, 1…
$ `SO 1`   <dbl> 1298, 1064, 1013, 1120, 1185, 1022, 1129, 1211, 1277, 1114, 1226, 990, 1091, 1054, 106…
$ BA   <dbl> 0.253, 0.263, 0.268, 0.270, 0.255, 0.258, 0.247, 0.264, 0.261, 0.260, 0.268, 0.260, 0.…
$ OBP   <dbl> 0.324, 0.339, 0.332, 0.352, 0.332, 0.329, 0.318, 0.339, 0.343, 0.331, 0.340, 0.319, 0.…
$ SLG   <dbl> 0.418, 0.405, 0.415, 0.454, 0.407, 0.411, 0.394, 0.417, 0.441, 0.416, 0.416, 0.400, 0.…
$ GDP   <dbl> 93, 142, 131, 137, 134, 139, 103, 140, 111, 131, 110, 153, 135, 128, 141, 128, 147, 14…
$ LOB   <dbl> 1173, 1223, 1160, 1210, 1209, 1086, 1131, 1198, 1147, 1159, 1214, 1080, 1090, 1134, 12…
$ salary   <dbl> 73516666, 96726166, 67101666, 121745999, 134809000, 96068500, 73558500, 81579166, 7520…
$ WAR   <dbl> 26.3, 42.4, 22.4, 50.8, 34.5, 34.5, 24.7, 26.0, 39.7, 33.4, 32.7, 19.4, 20.0, 43.8, 48…
```

(a) Every single variable in our data set. "League" is not a variable and Year and Team are combined into one variable. W, the team's Win Total for that season, is our response variable for model building.

```
```{r}
glimpse(MLBTeamStats1)
```

Rows: 480
Columns: 30
$ Year   <chr> "2009", "2009", "2009", "2009", "2009", "2009", "2009", "2009", "2009", "2009", "2009", …
$ Team   <chr> "ARI", "ATL", "BAL", "BOS", "CHC", "CHW", "CIN", "CLE", "COL", "DET", "FLA", "HOU", "KCR…
$ `RA/G` <dbl> 4.83, 3.96, 5.41, 4.54, 4.17, 4.52, 4.46, 5.34, 4.41, 4.57, 4.73, 4.75, 5.20, 4.70, 3.77…
$ DefEff <dbl> 0.683, 0.686, 0.680, 0.678, 0.698, 0.689, 0.704, 0.681, 0.688, 0.694, 0.686, 0.677, 0.67…
$ E   <dbl> 124, 96, 90, 82, 105, 113, 89, 97, 87, 88, 106, 78, 117, 85, 83, 98, 76, 97, 86, 105, 76…
$ DP   <dbl> 133, 159, 151, 121, 144, 158, 161, 170, 146, 164, 129, 161, 159, 174, 134, 149, 135, 134…
$ W   <dbl> 70, 86, 64, 95, 83, 79, 78, 65, 92, 86, 87, 74, 65, 97, 95, 80, 87, 70, 103, 75, 93, 62,…
$ L   <dbl> 92, 76, 98, 67, 78, 83, 84, 97, 70, 77, 75, 88, 97, 65, 67, 82, 76, 92, 59, 87, 69, 99, …
$ `W-L%` <dbl> 0.432, 0.531, 0.395, 0.586, 0.516, 0.488, 0.481, 0.401, 0.568, 0.528, 0.537, 0.457, 0.40…
$ ERA   <dbl> 4.42, 3.57, 5.15, 4.35, 3.84, 4.14, 4.18, 5.06, 4.22, 4.29, 4.29, 4.54, 4.83, 4.45, 3.41…
$ tSho   <dbl> 12, 10, 3, 11, 8, 11, 12, 6, 7, 9, 5, 10, 9, 13, 9, 8, 7, 12, 8, 10, 9, 7, 9, 10, 18, 11…
$ H   <dbl> 1470, 1399, 1633, 1494, 1329, 1438, 1420, 1570, 1427, 1449, 1425, 1521, 1486, 1513, 1265…
$ ER   <dbl> 711, 581, 817, 695, 616, 663, 677, 806, 675, 690, 690, 722, 765, 715, 558, 770, 726, 705…
$ HR   <dbl> 168, 119, 218, 167, 160, 169, 188, 183, 141, 182, 160, 176, 166, 180, 127, 207, 185, 158…
$ BB   <dbl> 525, 530, 546, 530, 586, 507, 577, 598, 528, 594, 601, 546, 600, 523, 584, 607, 466, 616…
$ SO   <dbl> 1158, 1232, 933, 1230, 1272, 1119, 1069, 986, 1154, 1102, 1248, 1144, 1153, 1062, 1272, …
$ `R/G`   <dbl> 4.44, 4.54, 4.57, 5.38, 4.39, 4.47, 4.15, 4.77, 4.96, 4.56, 4.77, 3.97, 4.23, 5.45, 4.81…
$ R   <dbl> 720, 735, 741, 872, 707, 724, 673, 773, 804, 743, 772, 643, 686, 883, 780, 785, 817, 671…
$ H1   <dbl> 1408, 1459, 1508, 1495, 1398, 1410, 1349, 1468, 1408, 1443, 1493, 1415, 1432, 1604, 1511…
$ RBI   <dbl> 686, 700, 708, 822, 678, 695, 637, 730, 760, 718, 727, 616, 657, 841, 739, 757, 770, 631…
$ SB   <dbl> 102, 58, 76, 126, 56, 113, 96, 84, 106, 72, 75, 113, 88, 148, 116, 68, 85, 122, 111, 133…
$ `SO 1` <dbl> 1298, 1064, 1013, 1120, 1185, 1022, 1129, 1211, 1277, 1114, 1226, 990, 1091, 1054, 1068,…
$ BA   <dbl> 0.253, 0.263, 0.268, 0.270, 0.255, 0.258, 0.247, 0.264, 0.261, 0.260, 0.268, 0.260, 0.25…
$ OBP   <dbl> 0.324, 0.339, 0.332, 0.352, 0.332, 0.329, 0.318, 0.339, 0.343, 0.331, 0.340, 0.319, 0.31…
$ SLG   <dbl> 0.418, 0.405, 0.415, 0.454, 0.407, 0.411, 0.394, 0.417, 0.441, 0.416, 0.416, 0.400, 0.40…
$ GDP   <dbl> 93, 142, 131, 137, 134, 139, 103, 140, 111, 131, 110, 153, 135, 128, 141, 128, 147, 144,…
$ LOB   <dbl> 1173, 1223, 1160, 1210, 1209, 1086, 1131, 1198, 1147, 1159, 1214, 1080, 1090, 1134, 1224…
$ salary <dbl> 73516666, 96726166, 67101666, 121745999, 134809000, 96068500, 73558500, 81579166, 752010…
$ WAR   <dbl> 26.3, 42.4, 22.4, 50.8, 34.5, 34.5, 24.7, 26.0, 39.7, 33.4, 32.7, 19.4, 20.0, 43.8, 48.8…
$ League <fct> 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1…
```

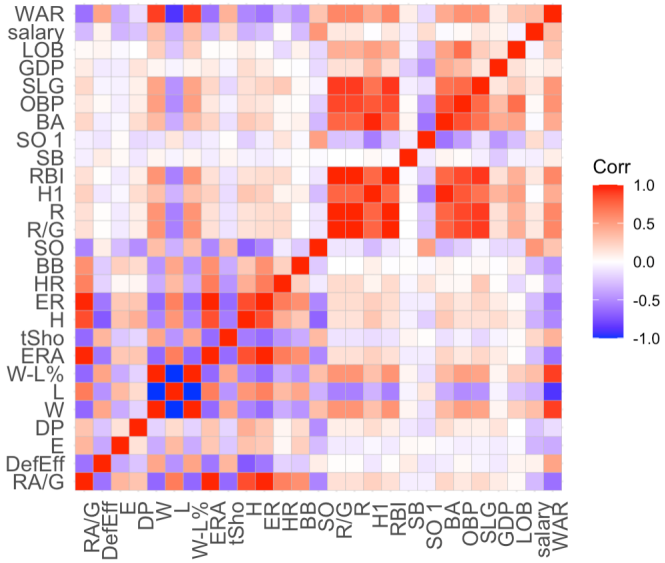
(b) Now Year and Team are separated into two separate variables and we have a League variable to represent whether a team plays for the American League or the National League.



(Appendix Fig. 2) Distribution of Wins



(Appendix Fig. 3) Diagnostics of the "Casual" Model



(Appendix Fig. 4) The correlation of every single variable.

(Appendix Fig. 5) Comparing two summary plots

```
Call:
lm(formula = W ~ WAR + League + ERA + E + H1 + BA + H + R + salary +
    OBP:salary, data = MLBTeamStats1)

Residuals:
     Min       1Q   Median       3Q      Max
-11.8249  -2.7397  -0.1435   2.4860  12.4066

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.755e+01  8.638e+00   3.189 0.001522 **
WAR          3.446e-01  7.082e-02   4.866 1.56e-06 ***
League1      1.421e+00  4.928e-01   2.884 0.004110 **
ERA         -1.193e+01  1.170e+00 -10.198  < 2e-16 ***
E           -4.203e-02  1.314e-02  -3.199 0.001474 **
H1          -8.346e-02  1.323e-02  -6.310 6.48e-10 ***
BA           5.671e+02  9.097e+01   6.234 1.01e-09 ***
H            9.210e-03  3.858e-03   2.387 0.017360 *
R            7.375e-02  6.787e-03  10.866  < 2e-16 ***
salary       2.374e-07  6.756e-08   3.515 0.000483 ***
salary:OBP  -6.934e-07  2.065e-07  -3.358 0.000849 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.861 on 469 degrees of freedom
Multiple R-squared:  0.8868,    Adjusted R-squared:  0.8844
F-statistic: 367.4 on 10 and 469 DF,  p-value: < 2.2e-16
```

(a) Summary of our model (mod3)

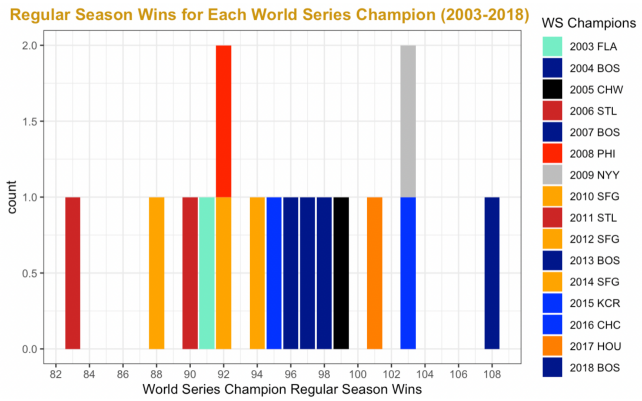
```
Call:
lm(formula = W ~ WAR + League + R + ERA + E + salary + H1 + BA +
    H + OBP:salary, data = MLBTeamStats1[-c(98, 158, 170), ])

Residuals:
     Min       1Q   Median       3Q      Max
-10.3792  -2.7882  -0.1251   2.5381  11.4302

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.345e+01  8.443e+00   2.778  0.00570 **
WAR          3.432e-01  6.912e-02   4.966 9.62e-07 ***
League1      1.364e+00  4.799e-01   2.842  0.00469 **
R            7.421e-02  6.642e-03  11.172  < 2e-16 ***
ERA         -1.220e+01  1.143e+00 -10.672  < 2e-16 ***
E           -4.026e-02  1.279e-02  -3.148  0.00175 **
salary       2.828e-07  6.636e-08   4.262 2.45e-05 ***
H1          -8.326e-02  1.289e-02  -6.459 2.66e-10 ***
BA           5.847e+02  8.861e+01   6.599 1.13e-10 ***
H            9.215e-03  3.772e-03   2.443  0.01495 *
salary:OBP  -8.357e-07  2.030e-07  -4.117 4.54e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.755 on 466 degrees of freedom
Multiple R-squared:  0.8929,    Adjusted R-squared:  0.8906
F-statistic: 388.4 on 10 and 466 DF,  p-value: < 2.2e-16
```

(b) Summary of our model after deleting possible influential points



(Appendix Fig. 6) Regular Season Wins of all World Series Champions from 2003-2018