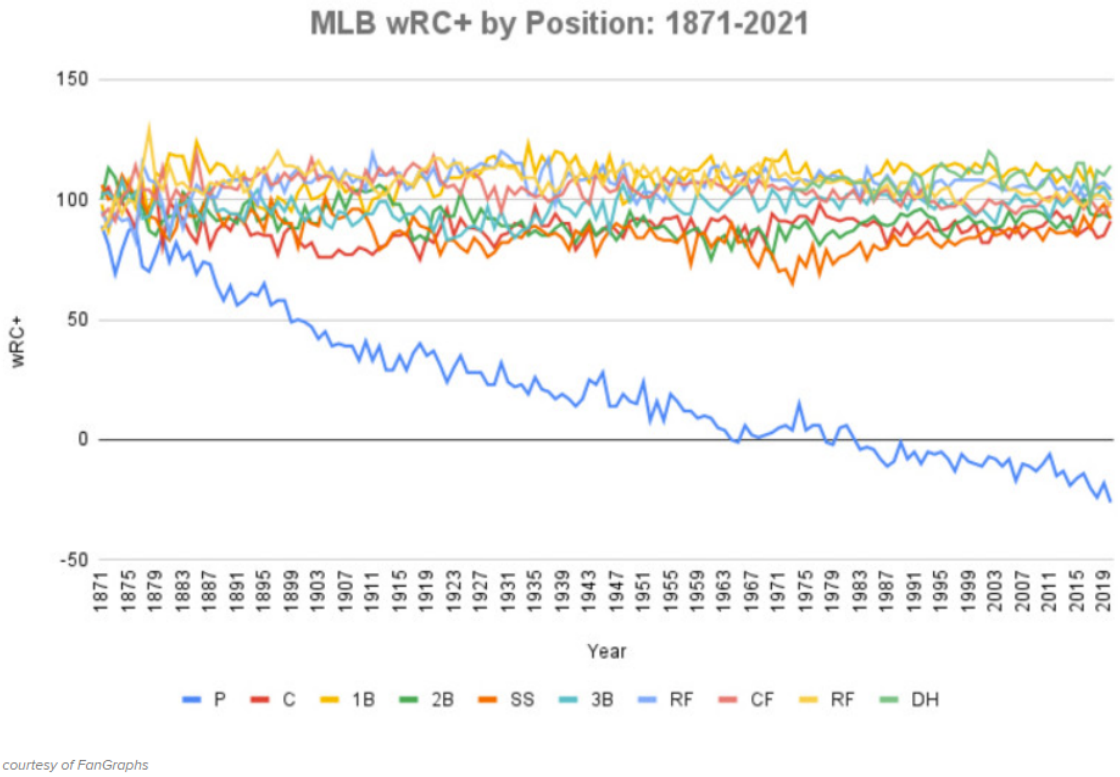
# Effects of Designated Hitter on Offense in Major League Baseball

## Introduction

Major League Baseball consists of two leagues – the American League (referred to as the “AL”) and the National League (“NL”). For decades, the leagues have both operated as a collective to make up MLB, each with a relatively equal number of teams, an equal geographical distribution of teams and the same set of rules. In almost every way, the leagues were identical, but there was one key differentiator: the presence of the designated hitter. In the National League, all nine players that played the field were also required to bat in the lineup, but the American League allowed teams to select a “designated hitter” to bat in place of the pitcher in the lineup. The AL opted to forgo pitcher’s hitting due to the historical inability for pitchers to hit successfully. The following graph taken from Bleacher Report shows the comparison of weighted runs created plus (wRC+), an all-encompassing offensive statistic, between positions (suffice to say a low wRC+ indicates a poor hitter).




*Data courtesy of FanGraphs*

In 2022, the leagues abandoned any individuality as the National League adopted the designated hitter (“DH”) themselves. The motives were clear – Major League Baseball wanted to remove the distinction between the leagues, whether the National League adopted the DH or the American League discarded it. The argument for adopting the DH in both leagues is the same reason the AL has had the DH since 1973 – replacing the pitcher with a capable hitter in the lineup would boost offense. By this argument, there should have been a significant difference in offensive production between the leagues while the National League did not have a designated hitter, and that difference should have disappeared in 2022.

For this project, I will be analyzing the difference in several offensive rate statistics between the leagues both when the National League did not have a designated hitter and when it had adopted the DH to test the assumption that the presence of the DH has a significant impact on offense. First, I will use Hotelling’s $T^2$ test to test for differences in the American and National Leagues without the DH in the NL, differences in leagues with the DH in the NL, and finally differences between the NL without the DH and the NL with the DH. If a difference is apparent, I will use simultaneous confidence intervals to determine where the difference exists.

## Data Description

The data for this project will be taken from the Lahman baseball database, perhaps the most reliable and widely used database for baseball statistics publicly available. From this, four different datasets will be extracted. The first two will be AL and NL batters from 2013 to 2021. This data will begin in 2013 as this was the year that the Houston Astros moved from the National League to the American League, creating an even 15 team split between leagues, making for more symmetric data analysis. Additionally, the year 2020 will be excluded from the data set due to a truncated season from the COVID-19 pandemic making for incomplete and unreliable data from this season. The next two data sets will be AL and NL batters in 2022. 2023 is excluded from this data set as this data is not publicly available at time of writing.

The offensive statistics (variables) we will be considering in this study will be batting average (“BA”), on base percentage (“OBP”) and slugging percentage (“SLG”). Each of these is a rate statistic, which we will be specifically using in place of counting statistics in order to maintain normality in the study. Further explanation of these statistics can be found at FanGraphs.com.

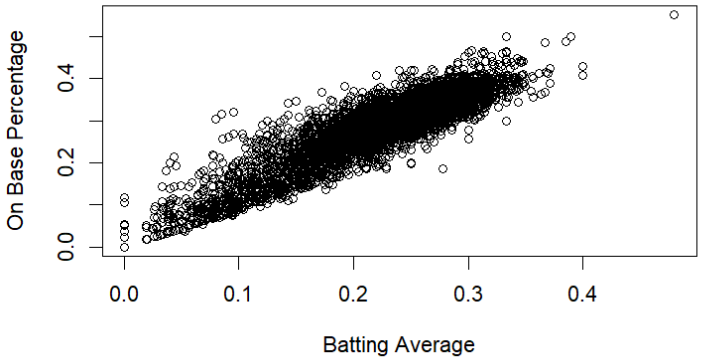
Although it would be beneficial to limit the player pool to those who were considered qualified batters by MLB’s definition (502 or more plate appearances in a year), this would remove all pitchers from our data set. However, we will limit the data set to only

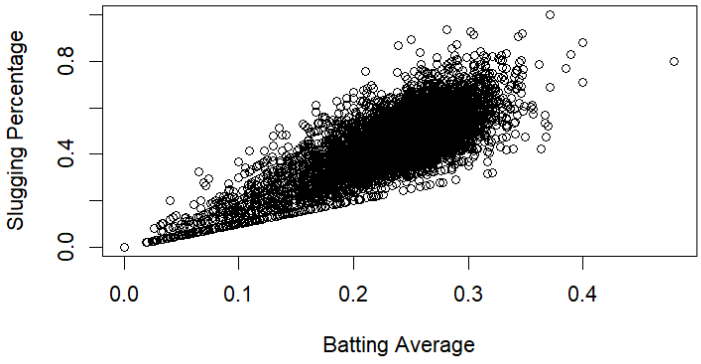
those who have 25 or more plate appearances in a year to eliminate the most extreme of outliers (for example, a player who came to bat one time and got a hit has a 1.000 batting average, where among qualified batters the highest batting average is almost always below .400).

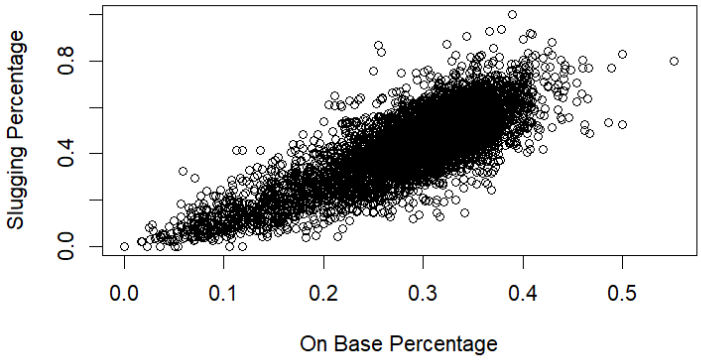
## Exploratory Data Analysis

The first step in data analysis for me was to plot each variable to examine the assumption of normality and identify any outliers.

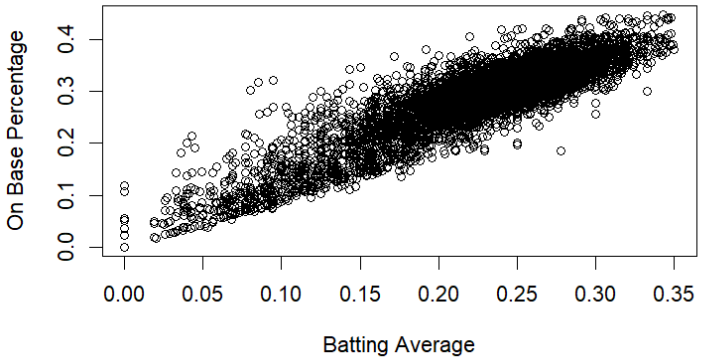
Firstly, I plotted each combination of variables as a scatterplot to identify any outliers and examine the plot for a roughly elliptical shape (which would indicate bivariate normality). The plots are shown here.

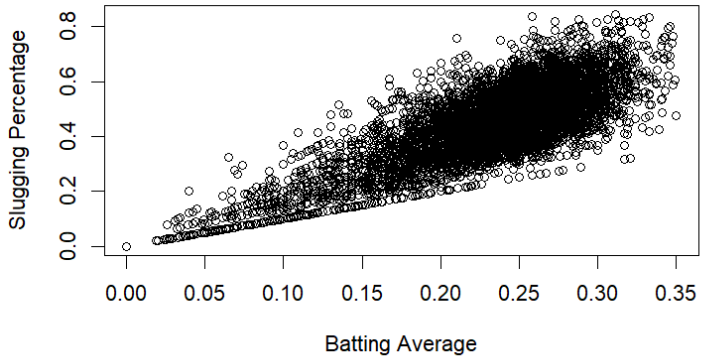


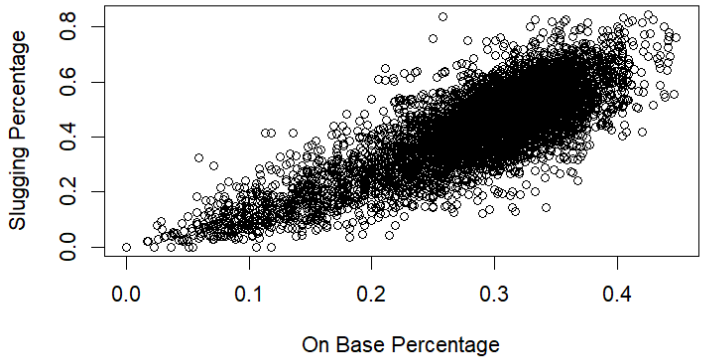






As can be seen in these plots, they all follow an elliptical pattern, but there are quite a few outliers. My method for removing these outliers was to visually inspect the plots and count the number of outliers. I then estimated the point at which the outliers began and selected these rows in the data (for example, all rows where the batters batting average was greater than 0.375). I then adjusted this number up and down until it matched the number of outliers I expected to be removed and did this for each variable. This is a crude method, however with very few outliers altogether that needed to be removed (most players needed to be included in the data set unless it was clear their production came only from a few number of plate appearances), it worked reasonably well. After removing the outliers, the plots now looked like this:






Each of these plots now follows a very elliptical pattern, and we can be confident that the data follow a multivariate normal distribution.

Another note on these plots is the relationship between each variable. Although not surprising, we can still highlight the fact that as one offensive statistic increases, the other is expected to increase as well, regardless of which one we look at. The correlation between batting average and on base percentage is 0.891, the correlation between batting average and slugging percentage is 0.816, and the correlation between on bae percentage and slugging percentage is 0.809, each indicating a very strong correlation.

Presented here are the summary statistics for each group with outliers removed:

**Means ($\bar{x}$)**

| | BA | OBP | SLG | $\bar{x}_i$ |
|---|---|---|---|---|
| AL (2013 – 2021) | 0.238 | 0.301 | 0.466 | $\bar{x}_1 = (0.238, 0.301, 0.466)$ |
| NL (2013 – 2021) | 0.217 | 0.274 | 0.410 | $\bar{x}_2 = (0.217, 0.274, 0.410)$ |
| AL (2022) | 0.226 | 0.292 | 0.433 | $\bar{x}_3 = (0.226, 0.292, 0.433)$ |

| | | | | |
|---|---|---|---|---|
| NL (2022) | 0.228 | 0.295 | 0.444 | $\bar{x}_4 = (0.228, 0.295, 0.444)$ |
| **Overall** | **0.227** | **0.287** | **0.435** | $\bar{x} = (0.227, 0.287, 0.435)$ |

**Standard Deviations (s)**

| | BA | OBP | SLG |
|---|---|---|---|
| AL (2013 – 2021) | 0.047 | 0.051 | 0.124 |
| NL (2013 – 2021) | 0.069 | 0.085 | 0.167 |
| AL (2022) | 0.049 | 0.051 | 0.116 |
| NL (2022) | 0.045 | 0.051 | 0.123 |
| **Overall** | **0.059** | **0.071** | **0.149** |

At a glance, the results from the summary statistics seem to align with the hypothesis that the National League would be worse at hitting from 2013 to 2021 than the American League, while the two leagues would be relatively equal in 2022. Indeed, from 2013 to 2021, the NL was quite lower than the AL in each offensive statistic. In 2022, the numbers are extremely close, however the AL did take the edge in batting average, while the NL lead in slugging percentage. This could be due to a random distribution or performance of players in this single year, or there may be some effect of the first year of the designated hitter, potentially (although I would wager that the difference here is due to randomness). Overall, however, the statistics between the leagues in 2022 are very close. Finally, comparing the National League in 2013 to 2021 to the NL in 2022, there also seems to be quite a difference. Just from the summary statistics, it seems that there is some effect of the designated hitter on offense.

Also of note is that, for each of the rate stats, the NL from 2013 to 2021 had the largest standard deviation. This is to be expected, as the presence of pitchers not only pulled the offensive mean down, but also created more separation between the rest of the hitters in the league, who sat around a league average level, and pitchers, who were far below this level.

Overall, the standard deviations seem to be close with a few differences. A look at the covariance matrices for each group will give us a clearer picture as to whether we can consider the variance to be equal for each group.

**Covariance for American League (2013-2021) ($s_1$)**

$$\begin{bmatrix} 0.002 & 0.002 & 0.004 \\ 0.002 & 0.003 & 0.005 \\ 0.004 & 0.005 & 0.015 \end{bmatrix}$$

**Covariance for National League (2013-2021) ($s_2$)**

$$\begin{bmatrix} 0.005 & 0.005 & 0.010 \\ 0.005 & 0.007 & 0.012 \\ 0.010 & 0.012 & 0.028 \end{bmatrix}$$

**Covariance for American League (2022) ($s_3$)**

$$\begin{bmatrix} 0.002 & 0.002 & 0.004 \\ 0.002 & 0.003 & 0.004 \\ 0.004 & 0.004 & 0.013 \end{bmatrix}$$

**Covariance for National League (2022) ($s_4$)**

$$\begin{bmatrix} 0.002 & 0.002 & 0.004 \\ 0.002 & 0.003 & 0.004 \\ 0.004 & 0.004 & 0.015 \end{bmatrix}$$

Performing Box's M test with these covariance matrices gives a p-value of approximately 0, indicating that there is significant evidence that at least one of these covariance matrices is different from the others. Indeed, by inspection that seems to be the case, as each value in $s_2$ is larger than each of the other covariance matrices. However, this analysis will not be to compare all mean vectors. In fact, we will be performing 3 different comparisons, and therefore will need to identify which pairs of covariance matrices can be considered equal and which ones cannot be.

The three different comparisons we will be making will be between AL and NL in the years 2013 to 2021, AL and NL in 2022, and between NL in 2013 to 2021 and NL in 2022. The results in Box's M test when comparing $s_1$ and $s_2$, $s_3$ and $s_4$, and $s_2$ and $s_4$ (or the variance-covariances matrices for each of these comparisons) confirms what our initial belief is, which is that the variance for the NL from 2013 to 2021 ($s_2$) is significantly different from each of the other matrices compared to, however the variance between the AL and NL in 2022 was not significantly different, and the comparison test between these two can be conducted assuming the variance is equal. The specific results are illustrated below.

**Box's M Test**

| Comparison | Chi-Square Approximation | Degrees of Freedom | P-Value |
|---|---|---|---|
| $s_1$ vs. $s_2$ | 655.03 | 6 | < 2.2e-16 |
| $s_3$ vs. $s_4$ | 7.865 | 6 | 0.2482 |
| $s_2$ vs. $s_4$ | 118.68 | 6 | < 2.2e-16 |

## Multivariate Analysis

The first step of this multivariate analysis is to analyze the difference between the American League and the National League from 2013 to 2021. This will be done using Hotelling's $T^2$ test. In this case, each data set is assumed to be multivariate normal, however, as detailed previously, $\Sigma_1 \neq \Sigma_2$, so this condition must be considered. Because the sample size is so large, we can use the chi-square approximation for this test. The test will be performed at the 0.05 level of significance.

Aside from the assumption of normality, we also must consider the assumption of independence. In this case, this assumption should hold true, as the offensive success or failure of one batter should not impact the success or failure of another.

For $\Sigma_1 \neq \Sigma_2$ and $n_1$, $n_2$ large, the $T^2$ test statistic is as follows:

$$T^2 = (\bar{x}_1 - \bar{x}_2)^T \left(\frac{1}{n_1}s_1 + \frac{1}{n_2}s_2\right)^{-1} (\bar{x}_1 - \bar{x}_2) \sim \chi_p^2$$

In this case, $T^2 = 226.3095$, which is greater than $X_3^2 = 7.185$, so we can indeed reject the null hypothesis that the mean vectors of the American League from 2013 to 2021 and the National League over the same time frame are equal.

We now repeat the process for the American League and National League in 2022, however in this case, we can assume the covariance matrices are equal and use an approximate F distribution for the test. Our test statistic in this case is as follows:

$$T^2 = (\bar{x}_1 - \bar{x}_2)^T \left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s_p\right)^{-1} (\bar{x}_1 - \bar{x}_2) \sim \left(\frac{n_1 + n_2 - 2}{n_1 + n_2 - p - 1}\right) F_{p, n_1+n_2-p-1}$$

$$where\ s_p = \frac{(n_1 - 1)s_1 + (n_2 - 1)s_2}{n_1 + n_2 - 2}$$

Here, our $T^2$ value is 1.462, which is smaller than the critical value of 2.627, indicating that, at the 95% level of confidence, there is not a significant difference between the offensive performance of AL batters and NL batters in 2022.

Finally, we analyze the difference between NL batters from 2013 to 2021 and NL batters in 2022. This is again done with the assumption that $\Sigma_1 \neq \Sigma_2$ and $n_1$ and $n_2$ are large to use the chi-square approximation. Using the same formula as presented above, we get a $T^2$ value of 46.810, which is again larger than our critical value of 7.815. This again confirms our hypothesis that NL batting in 2022 is significantly different from NL batting from 2013 to 2021.

We would now like to further explore the differences between the NL from 2013 to 2021 and the groups it was compared to in order to identify where the significant differences come from. To do this, we will use simultaneous confidence intervals. For each case, we are unable to assume that $\Sigma_1 = \Sigma_2$, and so we will have to use the appropriate formula to incorporate the chi-square approximation. That formula is given by:

$$\bar{x}_{1i} - \bar{x}_{2i} \pm \sqrt{\chi_p^2}\sqrt{c_{ii}}$$

$$where\ c = \frac{1}{n_1}s_1 + \frac{1}{n_2}s_2$$

Computing this for the National League and American League comparison from 2013 to 2021, we get the following confidence intervals (for alpha = 0.05):

**Confidence Intervals for AL (2013 – 2021) vs. NL (2013 – 2021)**

| Variable | Lower Bound | Upper Bound |
|---|---|---|
| Batting Average | 0.009 | 0.031 |
| On Base Percentage | 0.013 | 0.041 |
| Slugging Percentage | 0.029 | 0.083 |

This indicates that each variable was significantly different when comparing the two leagues Because we set $x_1$ to be the American League, we can also tell that the direction of the difference suggests that the American League was significantly *better* than the National League in terms of offensive output over the years 2013 to 2021.

Repeating this analysis with the National League over our different time periods, we get the following table of confidence intervals:

**Confidence Intervals for NL (2013 – 2021) vs. NL (2022)**

| Variable | Lower Bound | Upper Bound |
|---|---|---|
| Batting Average | -0.019 | -0.003 |
| On Base Percentage | -0.030 | -0.112 |
| Slugging Percentage | -0.055 | -0.013 |

Again we see that each variable was significantly different, however this time having set NL (2013 – 2021) as $x_1$, we now interpret the direction of the difference as the NL in 2022 having been significantly better than the NL from 2013 – 2021, which is the direction we expected.

# Conclusion

Coming into this study, we expected that the National League from 2013 – 2021 would perform worse than the AL over the same period as well as the NL once it had adopted the DH in 2022 due to the impact of pitchers hitting. Indeed, these findings confirm that hypothesis, as the Hotelling $T^2$ test proved significance in each case. Additionally, if the presence of a DH truly does make an impact, then the leagues should have remained equal in 2022 and no significance would be found in the $T^2$ when comparing the leagues in 2022, and this is precisely what we found. Overall, it seems the conclusion that can be drawn is that our hypothesis was correct, and the designated hitter does significantly impact offensive production.

Even with the evidence provided, there are other possible reasons we could have seen these differences. The most likely other explanation would be that National League over this period just happened to have a poor distribution of players, the other league had a more productive group over the same period, and then in 2022 the NL improved its player pool. While possible, I would still claim that it's more probable that the designated hitter was the cause of this shift rather than random chance. Another possible explanation is that other changes to the game impacted the leagues in different ways. For example, not only was the designated hitter introduced in 2022, but a new rule was created that allowed each team to start an inning with a runner on 2$^{nd}$ base in extra innings. This change could have impacted the leagues differently, however seeing as the rule was identical between leagues, I find it unlikely that this would have had a disproportionate impact on the leagues.

Overall, I believe our findings are conclusive, and the addition of the designated hitter in the National League has boosted offense for the players therein. This topic could be studied further in the future with more research involving a larger pool of data and more sophisticated statistics to define the impact of the designated hitter more precisely, but for now I believe we have proven the effect exists.

## References

Rymer, Z. D. (2021, May 13). *MLB pitchers' historically awful hitting should seal the fate of a universal DH*. Bleacher Report. https://bleacherreport.com/articles/2943276-mlb-pitchers-historically-awful-hitting-should-seal-the-fate-of-a-universal-dh

*Seanlahman.com*. SeanLahman.com. (n.d.). http://seanlahman.com/

Slowinski, P. (2013, June 25). *Fangraphs library stat glossary*. Sabermetrics Library. https://library.fangraphs.com/fangraphs-library-glossary/

# Appendix

```
library(biotools)

## Warning: package 'biotools' was built under R version 4.1.3

## Loading required package: MASS

## Warning: package 'MASS' was built under R version 4.1.3

## ---
## biotools version 4.2

###Read in data
AL1 <- read.csv("C:/School/Multivariate/FinalProject/AL_Data_2013_2021.csv",
fileEncoding="UTF-8-BOM"); head(AL1)
```

```
##    playerID yearID teamID lgID HR SB RBI  PA    BA   OBP   SLG
## 1 casilal01   2013    BAL   AL  1  9  10 125 0.214 0.264 0.348
## 2 davisch02   2013    BAL   AL 53  4 138 673 0.286 0.370 0.798
## 3 dickech01   2013    BAL   AL  4  5  13 109 0.238 0.266 0.486
## 4 flahery01   2013    BAL   AL 10  2  27 271 0.224 0.292 0.476
## 5 hardyjj01   2013    BAL   AL 25  2  76 644 0.263 0.304 0.519
## 6 jonesad01   2013    BAL   AL 33 14 108 689 0.285 0.318 0.599
```

```
NL1 <- read.csv("C:/School/Multivariate/FinalProject/NL_Data_2013_2021.csv",
fileEncoding="UTF-8-BOM"); head(NL1)
```

```
##    playerID yearID teamID lgID HR SB RBI  PA    BA   OBP   SLG
## 1 bloomwi01   2013    ARI   NL  0  0  14 150 0.317 0.360 0.410
## 2 cahiltr01   2013    ARI   NL  0  0   5  53 0.082 0.075 0.184
## 3 campato01   2013    ARI   NL  0  8   0  54 0.261 0.370 0.326
## 4 chaveer01   2013    ARI   NL  9  1  44 253 0.281 0.332 0.588
## 5 corbipa01   2013    ARI   NL  0  0   3  73 0.123 0.151 0.246
## 6 davidma02   2013    ARI   NL  3  0  12  87 0.237 0.333 0.553
```

```
AL2 <- read.csv("C:/School/Multivariate/FinalProject/AL_Data_2022.csv", fileEncoding="UTF-8-
BOM"); head(AL2)
```

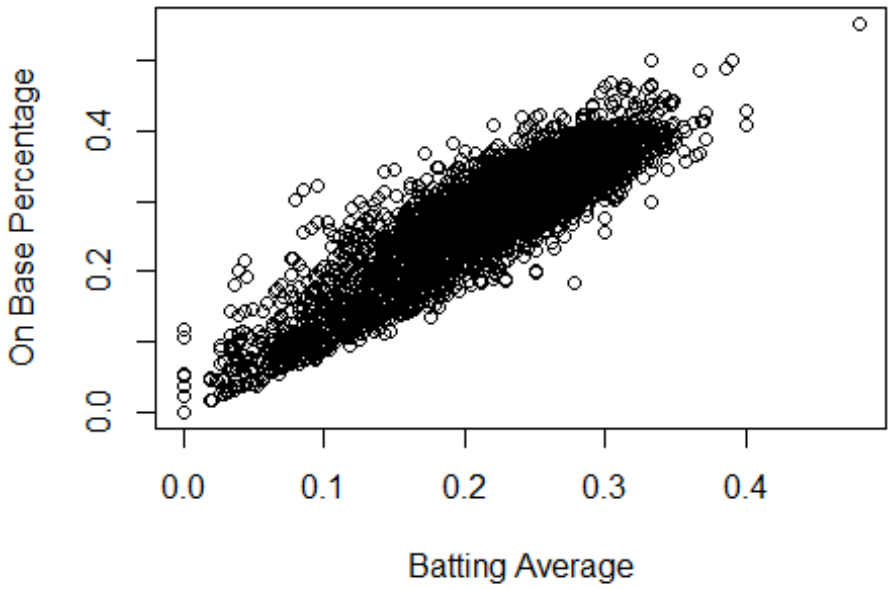
```
##    playerID yearID teamID lgID HR SB RBI  PA    BA   OBP   SLG
## 1 aguilje01   2022    BAL   AL  1  0   2  50 0.224 0.240 0.347
## 2 arauzjo01   2022    BAL   AL  1  0   4  29 0.179 0.207 0.321
## 3 bemboan01   2022    BAL   AL  1  0   1  59 0.115 0.203 0.269
## 4 chiriro01   2022    BAL   AL  4  1  22 220 0.179 0.264 0.354
## 5 gutieke01   2022    BAL   AL  0  1   3  32 0.143 0.250 0.214
## 6  haysau01   2022    BAL   AL 16  2  60 582 0.250 0.306 0.512
```

```
NL2 <- read.csv("C:/School/Multivariate/FinalProject/NL_Data_2022.csv", fileEncoding="UTF-8-
BOM"); head(NL2)
```

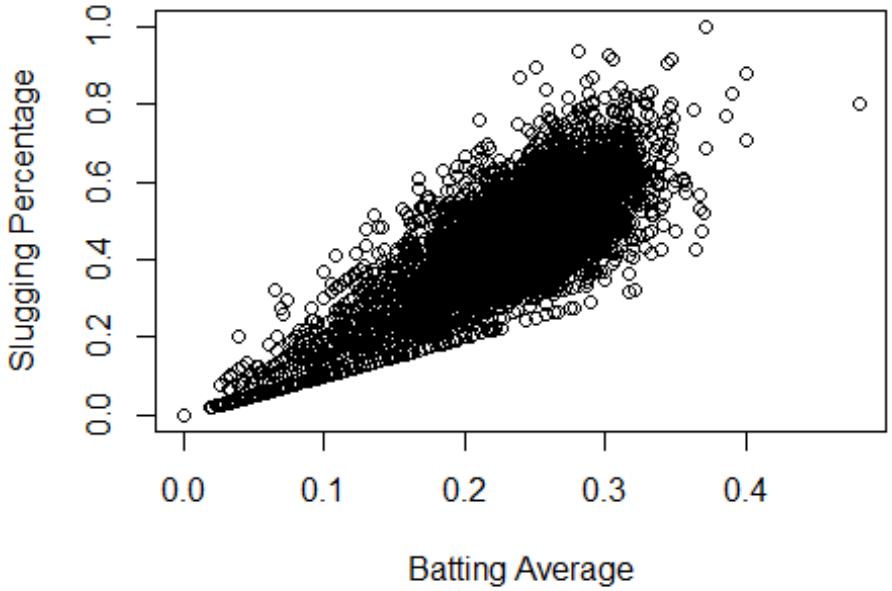
```
##    playerID yearID teamID lgID HR SB RBI  PA    BA   OBP   SLG
## 1 ahmedni01   2022    ARI   NL  3  0   7  54 0.231 0.259 0.538
## 2 alcanse01   2022    ARI   NL  1  0   7  56 0.189 0.196 0.396
```

```
## 3 alcanse01   2022    ARI   NL  5  1  19 130 0.265 0.315 0.538
## 4  beerse01   2022    ARI   NL  1  0   9 126 0.189 0.278 0.279
## 5 carroco02   2022    ARI   NL  4  2  14 115 0.260 0.330 0.644
## 6 garrest01   2022    ARI   NL  4  3  10  81 0.276 0.309 0.697
```

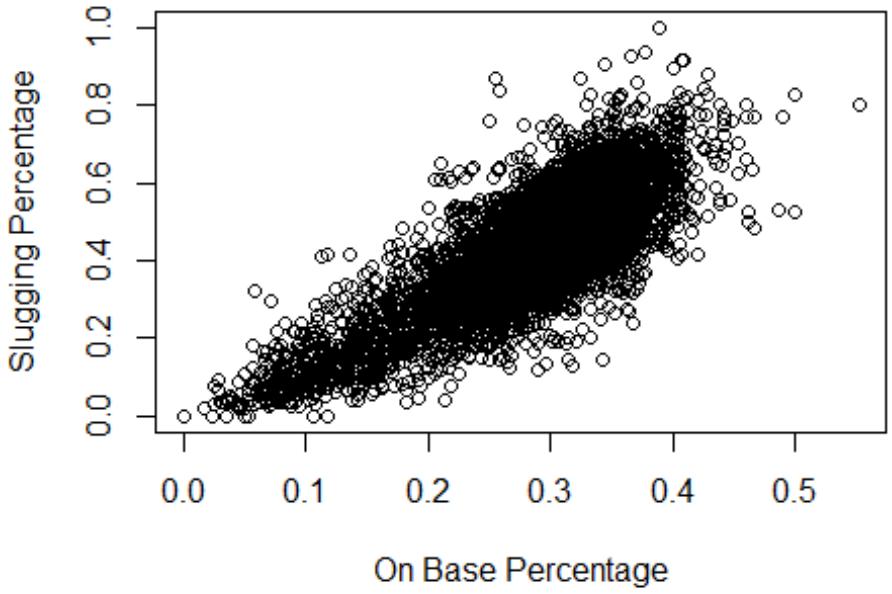
```
AL1 <- cbind(AL1[,c(-5, -6, -7)], Class = "AL1")
NL1 <- cbind(NL1[,c(-5, -6, -7)], Class = "NL1")
AL2 <- cbind(AL2[,c(-5, -6, -7)], Class = "AL2")
NL2 <- cbind(NL2[,c(-5, -6, -7)], Class = "NL2")
Full_Data <- rbind(AL1, AL2, NL1, NL2)

###Plot data to look for normality and outliers
plot(Full_Data[,6], Full_Data[,7], xlab = "Batting Average", ylab = "On Base Percentage")
```



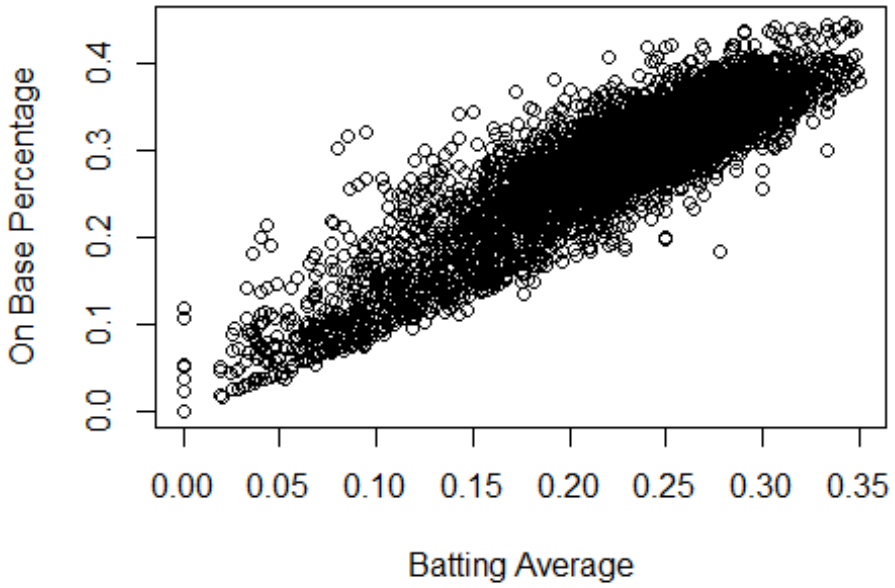

```
plot(Full_Data[,6], Full_Data[,8], xlab = "Batting Average", ylab = "Slugging Percentage")
```



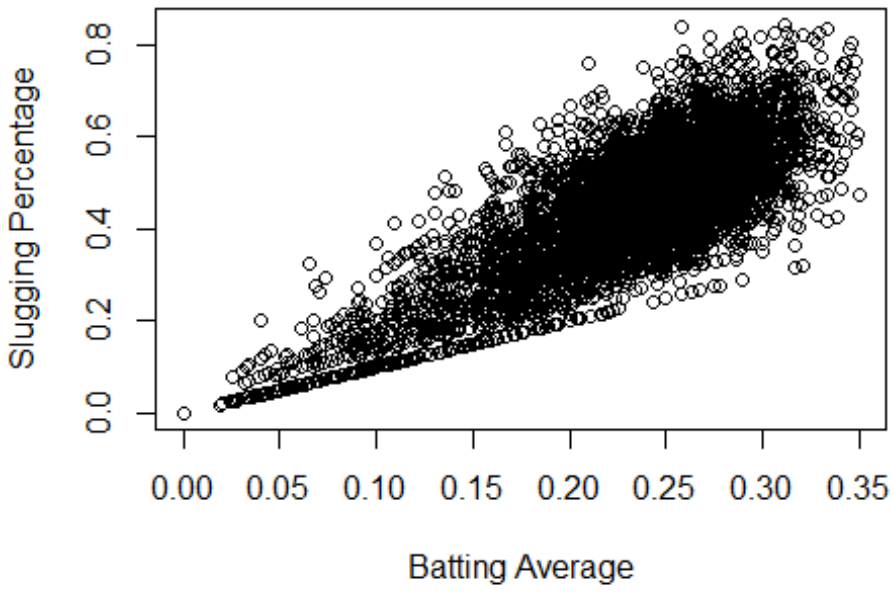

```
plot(Full_Data[,7], Full_Data[,8], xlab = "On Base Percentage", ylab = "Slugging Percentage")
```



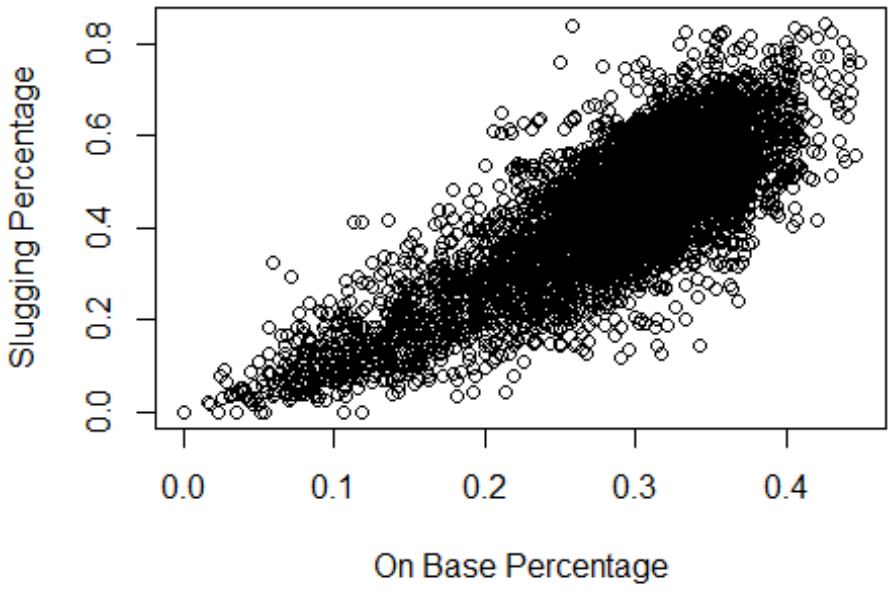
```
###Identify and remove outliers
outliers <- Full_Data[Full_Data$SLG > .85 | Full_Data$BA > .35 | Full_Data$OBP > .45,]
data <- Full_Data[Full_Data$BA <= .35 & Full_Data$OBP <= .45 & Full_Data$SLG <= .85,]

###Plot data without outliers
plot(data[,6], data[,7], xlab = "Batting Average", ylab = "On Base Percentage")
```




```
plot(data[,6], data[,8], xlab = "Batting Average", ylab = "Slugging Percentage")
```




```
plot(data[,7], data[,8], xlab = "On Base Percentage", ylab = "Slugging Percentage")
```



```r
###Identify correlation between variables
cor(data[,6], data[,7])
```

```
## [1] 0.890646
```

```r
cor(data[,6], data[,8])
```

```
## [1] 0.8155559
```

```r
cor(data[,7], data[,8])
```

```
## [1] 0.8085358
```

```r
###Split data into groups by league and year
groups <- split(data[,-9], data[,9])
AL1 <- groups$AL1
NL1 <- groups$NL1
AL2 <- groups$AL2
NL2 <- groups$NL2


###Analyze mean and standard deviation of each group and of the overall data
###and find variance-covariance matrices
mean(AL1$BA); mean(NL1$BA); mean(AL2$BA); mean(NL2$BA); mean(data$BA)
```

```
## [1] 0.2377007
```

```
## [1] 0.2172781
```

```
## [1] 0.2259432
```

```
## [1] 0.2284132
```

```
## [1] 0.2265332
```

```r
mean(AL1$OBP); mean(NL1$OBP); mean(AL2$OBP); mean(NL2$OBP); mean(data$OBP)
```

```
## [1] 0.3013526
```

```
## [1] 0.27443
```

```
## [1] 0.2924132

## [1] 0.294836

## [1] 0.2872783

mean(AL1$SLG); mean(NL1$SLG); mean(AL2$SLG); mean(NL2$SLG); mean(data$SLG)

## [1] 0.4656044

## [1] 0.4095768

## [1] 0.4327192

## [1] 0.4438044

## [1] 0.4351282

sd(AL1$BA); sd(NL1$BA); sd(AL2$BA); sd(NL2$BA); sd(data$BA)

## [1] 0.04692442

## [1] 0.06850526

## [1] 0.0487699

## [1] 0.04543764

## [1] 0.05934739

sd(AL1$OBP); sd(NL1$OBP); sd(AL2$OBP); sd(NL2$OBP); sd(data$OBP)

## [1] 0.05143307

## [1] 0.08498269

## [1] 0.05102639

## [1] 0.05142832

## [1] 0.07114026

sd(AL1$SLG); sd(NL1$SLG); sd(AL2$SLG); sd(NL2$SLG); sd(data$SLG)
```

```
## [1] 0.1235192

## [1] 0.167194

## [1] 0.1159542

## [1] 0.1229114

## [1] 0.1486543

sigma1 <- cov(AL1[,6:8])
sigma2 <- cov(NL1[,6:8])
sigma3 <- cov(AL2[,6:8])
sigma4 <- cov(NL2[,6:8])

###Test for equality of variance-covariance matrices
boxM(data[,6:8], data[,9])

##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  data[, 6:8]
## Chi-Sq (approx.) = 777.51, df = 18, p-value < 2.2e-16

###Test for equality of variance between AL1 and NL1
data_13_to_21 <- rbind(data[data$Class == "AL1", 6:9], data[data$Class == "NL1", 6:9])
boxM(data_13_to_21[,-4], data_13_to_21[,4])

##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  data_13_to_21[, -4]
## Chi-Sq (approx.) = 655.03, df = 6, p-value < 2.2e-16

###Test for equality of variance between AL2 and NL2
data_22 <- rbind(data[data$Class == "AL2", 6:9], data[data$Class == "NL2", 6:9])
boxM(data_22[,-4], data_22[,4])
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  data_22[, -4]
## Chi-Sq (approx.) = 7.8646, df = 6, p-value = 0.2482
```

***###Test for equality of variance between NL1 and NL2***

```
NL_data <- rbind(data[data$Class == "NL1", 6:9], data[data$Class == "NL2", 6:9])
boxM(NL_data[,-4], NL_data[,4])
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  NL_data[, -4]
## Chi-Sq (approx.) = 118.68, df = 6, p-value < 2.2e-16
```

***###T2 test for AL vs. NL (2013-2021)***

```
AL <- AL1[,6:8]
NL <- NL1[,6:8]
x1bar <- colMeans(AL); x1bar
```

```
##        BA       OBP       SLG
## 0.2377007 0.3013526 0.4656044
```

```
x2bar <- colMeans(NL); x2bar
```

```
##        BA       OBP       SLG
## 0.2172781 0.2744300 0.4095768
```

```
s1 <- cov(AL); s1
```

```
##             BA        OBP        SLG
## BA  0.002201901 0.002020794 0.004143084
## OBP 0.002020794 0.002645361 0.004518374
## SLG 0.004143084 0.004518374 0.015256983
```

```
s2 <- cov(NL); s2
```

```
##              BA         OBP         SLG
## BA  0.004692971 0.005310780 0.009832833
## OBP 0.005310780 0.007222057 0.012095545
## SLG 0.009832833 0.012095545 0.027953817
```

```
n1 <- nrow(AL); n1
```

```
## [1] 2419
```

```
n2 <- nrow(NL); n2
```

```
## [1] 2963
```

```
p <- ncol(AL); p
```

```
## [1] 3
```

```
alpha <- .05
```

```
#Find test statistic, crit value and compare
t2 <- t(x1bar-x2bar)%*%solve(((1/n1)*s1)+((1/n2)*s2))%*%(x1bar-x2bar); t2
```

```
##          [,1]
## [1,] 226.3095
```

```
crit_val <- qchisq(alpha, p, lower.tail = F); crit_val
```

```
## [1] 7.814728
```

```
t2 > crit_val
```

```
##      [,1]
## [1,] TRUE
```

```
###T2 test for AL vs. NL (2022)
AL <- AL2[,6:8]
NL <- NL2[,6:8]
x1bar <- colMeans(AL); x1bar
```

```
##        BA       OBP       SLG
## 0.2259432 0.2924132 0.4327192
```

```
x2bar <- colMeans(NL); x2bar
```

```
##        BA       OBP       SLG
## 0.2284132 0.2948360 0.4438044
```

```
s1 <- cov(AL); s1
```

```
##             BA        OBP        SLG
## BA  0.002378503 0.002052223 0.004134797
## OBP 0.002052223 0.002603693 0.003991208
## SLG 0.004134797 0.003991208 0.013445380
```

```
s2 <- cov(NL); s2
```

```
##             BA        OBP        SLG
## BA  0.002064579 0.001951856 0.004111103
## OBP 0.001951856 0.002644872 0.004400351
## SLG 0.004111103 0.004400351 0.015107221
```

```
n1 <- nrow(AL); n1
```

```
## [1] 317
```

```
n2 <- nrow(NL); n2
```

```
## [1] 317
```

```
p <- ncol(AL); p
```

```
## [1] 3
```

```
alpha <- .05

#Find test statistic, crit value and compare
sp <- ((n1-1)*s1+(n2-1)*s2)/(n1+n2-2)
t2 <- t(x1bar-x2bar)%*%solve((1/n1+1/n2)*sp)%*%(x1bar-x2bar); t2
```

```
##          [,1]
## [1,] 1.461712
```

```
crit_val <- ((n1+n2-2)/(n1+n2-p-1))*qf(alpha, p, (n1+n2-p-1), lower.tail = F); crit_val
```

```
## [1] 2.627359
```

```
t2 > crit_val
```

```
##       [,1]
## [1,] FALSE
```

```
###T2 test for NL Comparison
NL_1 <- NL1[,6:8]
NL_2 <- NL2[,6:8]
x1bar <- colMeans(NL_1); x1bar
```

```
##        BA       OBP       SLG
## 0.2172781 0.2744300 0.4095768
```

```
x2bar <- colMeans(NL_2); x2bar
```

```
##        BA       OBP       SLG
## 0.2284132 0.2948360 0.4438044
```

```
s1 <- cov(NL_1); s1
```

```
##             BA        OBP        SLG
## BA  0.004692971 0.005310780 0.009832833
## OBP 0.005310780 0.007222057 0.012095545
## SLG 0.009832833 0.012095545 0.027953817
```

```
s2 <- cov(NL_2); s2
```

```
##             BA        OBP        SLG
## BA  0.002064579 0.001951856 0.004111103
## OBP 0.001951856 0.002644872 0.004400351
## SLG 0.004111103 0.004400351 0.015107221
```

```
n1 <- nrow(NL_1); n1
```

```
## [1] 2963

n2 <- nrow(NL_2); n2

## [1] 317

p <- ncol(NL_1); p

## [1] 3

alpha <- .05


#Find test statistic, crit value and compare
t2 <- t(x1bar-x2bar)%*%solve(((1/n1)*s1)+((1/n2)*s2))%*%(x1bar-x2bar); t2

##          [,1]
## [1,] 46.81029

crit_val <- qchisq(alpha, p, lower.tail = F); crit_val

## [1] 7.814728

t2 > crit_val

##      [,1]
## [1,] TRUE

###Simultaneous confidence intervals for NL1 vs. AL1
AL <- AL1[,6:8]
NL <- NL1[,6:8]
x1bar <- colMeans(AL); x1bar

##        BA       OBP       SLG
## 0.2377007 0.3013526 0.4656044

x2bar <- colMeans(NL); x2bar

##        BA       OBP       SLG
## 0.2172781 0.2744300 0.4095768
```

```r
s1 <- cov(AL); s1
```

```
##             BA        OBP        SLG
## BA  0.002201901 0.002020794 0.004143084
## OBP 0.002020794 0.002645361 0.004518374
## SLG 0.004143084 0.004518374 0.015256983
```

```r
s2 <- cov(NL); s2
```

```
##             BA        OBP        SLG
## BA  0.004692971 0.005310780 0.009832833
## OBP 0.005310780 0.007222057 0.012095545
## SLG 0.009832833 0.012095545 0.027953817
```

```r
p <- ncol(AL); p
```

```
## [1] 3
```

```r
alpha <- .05

CI <- matrix(nrow = p, ncol = 2)
chi <- qchisq(alpha, p, lower.tail = F)
c <- 1/n1*s1+1/n2*s2

for (i in 1:p){
  CI[i, 1] <- (x1bar[i]-x2bar[i])-(sqrt(chi)*sqrt(c[i,i]))
  CI[i, 2] <- (x1bar[i]-x2bar[i])+(sqrt(chi)*sqrt(c[i,i]))
}; CI
```

```
##             [,1]       [,2]
## [1,] 0.009399941 0.03144527
## [2,] 0.013320592 0.04052472
## [3,] 0.029020922 0.08303428
```

```r
###Simultaneous confidence intervals for NL1 vs. AL1
NL_1 <- NL1[,6:8]
```

```r
NL_2 <- NL2[,6:8]
x1bar <- colMeans(NL_1); x1bar
```

```
##        BA       OBP       SLG
## 0.2172781 0.2744300 0.4095768
```

```r
x2bar <- colMeans(NL_2); x2bar
```

```
##        BA       OBP       SLG
## 0.2284132 0.2948360 0.4438044
```

```r
s1 <- cov(NL_1); s1
```

```
##             BA        OBP        SLG
## BA  0.004692971 0.005310780 0.009832833
## OBP 0.005310780 0.007222057 0.012095545
## SLG 0.009832833 0.012095545 0.027953817
```

```r
s2 <- cov(NL_2); s2
```

```
##             BA        OBP        SLG
## BA  0.002064579 0.001951856 0.004111103
## OBP 0.001951856 0.002644872 0.004400351
## SLG 0.004111103 0.004400351 0.015107221
```

```r
p <- ncol(NL_1); p
```

```
## [1] 3
```

```r
alpha <- .05

CI <- matrix(nrow = p, ncol = 2)
chi <- qchisq(alpha, p, lower.tail = F)
c <- 1/n1*s1+1/n2*s2

for (i in 1:p){
  CI[i, 1] <- (x1bar[i]-x2bar[i])-(sqrt(chi)*sqrt(c[i,i]))
```

```
  CI[i, 2] <- (x1bar[i]-x2bar[i])+(sqrt(chi)*sqrt(c[i,i]))
}; CI

##             [,1]         [,2]
## [1,] -0.01908963 -0.003180676
## [2,] -0.02958474 -0.011227241
## [3,] -0.05534994 -0.013105331
```