**Capstone Project Draft 1**

*Clutch Hitting in the Major Leagues: A Myth or Not?*

*Triet Tran*

**Introduction**

Baseball is a game of numbers. It has a unique ability to document nearly every action that takes place such as batting average, slugging percentage, and earned run average. As a result, statistics are prevalent in baseball because they are used primarily as predictors. Over a hundred years of statistics have shown us statistical norms as well as regression indicators within the game of baseball.

However, there is still an aspect shrouded in mystery: the fabled “clutch hitting” ability of a batter. In baseball, a “clutch” hitter is a player who often come up with the “big” hit. The big hit is typically a game-deciding hit, or any hit with a significant impact late in a game. These big hits occur in “clutch” situations such as two outs, late innings or when the team is losing. A clutch hitter is held in a position of high regard and responsibility because he is acknowledged as the "go-to guy" for the team, and his aptitude in pressure situations are celebrated by both players and fans. This project endeavors to look into the statistical validity of “clutch” hitting, and if such ability exists, the extent to which clutch hitting has on a baseball team’s performance.

**Data sources**

I used the data provided by Baseball Reference and ESPN which can be extracted into a .csv file. Surprisingly, clutch hitting statistics are not organized neatly and scattered all over these tow databases. Therefore, I optimized an algorithm to capture these bits of data (taking advantage of the repeated patterns of the URL links where the information is stored on ESPN). Afterwards, I compiled them into a collective R table so the analysis process would be expedited. Data from the year 2002 to 2017 were used for this project.

**Methods**

“Clutch hitting” statistics are useful to the extent that they can be rather descriptive of what happened during a baseball game or a season. However, many experts do not believe that “clutch hitting” could serve as a true judge of talent or as an method of prediction. It is argued that a “clutch” batter would have to be a bad player in normal situations, but happen to be better than the best players in high leverage situations time and time again.

The first priority of this project is to clearly define “clutch hitting” in order to set the bounds and parameters for subsequent statistical analysis. A player's batting average with RISP (Runners in Scoring Position) may be a simple example of "hitting in the clutch.” In contrast, a player’s batting average with no runner on base could be an example of “non-clutch” situation or the “control” of this experiment. This disparity is parallel with the two-stage process of run scoring: one puts runners on base (non-clutch) and then advance them to home (clutch situations). Since runs are the most important attributes in a baseball game (the team with more runs would win), teams are naturally interested in “clutch hitting:” the ability to score runners who are in scoring position.

**Analysis**

With this definition, the project aims to answer these following questions:

1) Does “clutch” hitting really exist?

There are 3,684, 169 at-bats from 2002 to 2017. However, it should be noted from Figure 1 that a large number of batters have less than 10 at-bats. This is because the data includes the at-bats of American League pitchers or relievers who do not have may opportunities to hit. These players account for roughly half the players of Major League Baseball (MLB). Therefore, I set the cut-off at the median (50th percentile) to omit batters with too few at-bats because small sample sizes often lead to over-inflated statistics. I used the median instead of because the data is heavily right-skewed.

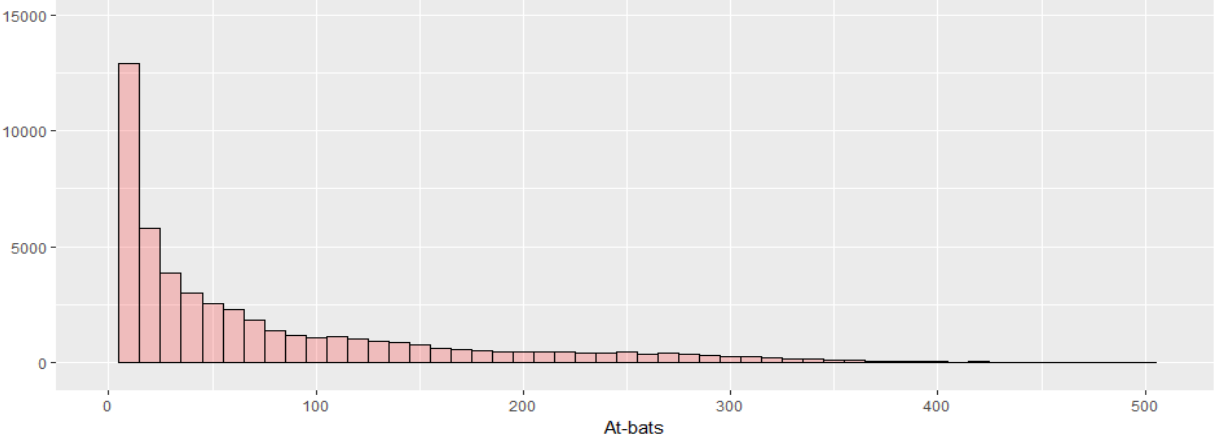


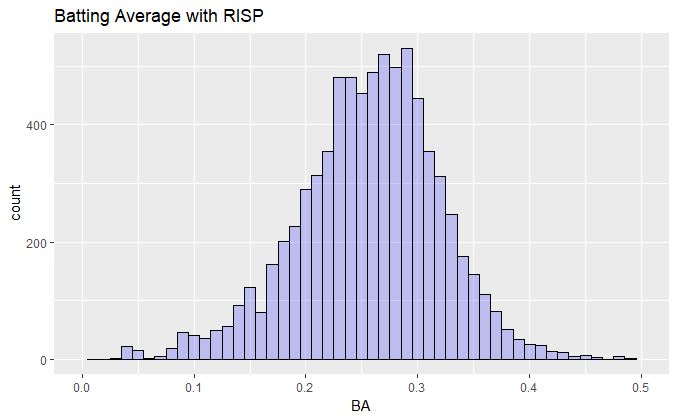
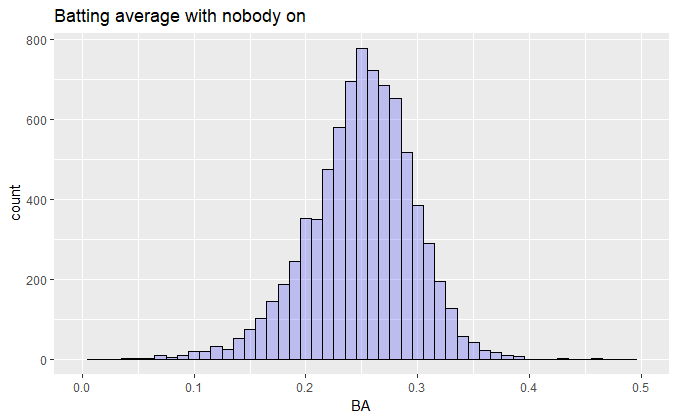
Figure 1 Total at-bats distribution

This leaves me 3,456,690 at-bats to analyze. I categorized these at-bats in the data into five situational hitting sub-groups:

* No runner on base (41.4%)
* Runner(s) on base (31.2%)
* Runner(s) in scoring position (17.6%)
* Runner(s) in scoring position and two outs (8.1%)
* With the bases loaded (1.6%)

The control, or “non-clutch” situations, consists of the at-bats where there is no runner on base. The clutch situations include the other four sub-groups listed above.

I plotted the distribution of batting averages when there is no runner on base versus when there is runner in scoring position (including with zero out, one out and two outs):



Interestingly, a higher batting average is observed when there is runner in scoring position (.257) versus when there is no runner on base (.250), p-value<0.01. However, a wider spread is observed for batting average with runner in scoring position. Why would these players have a statistically significantly higher batting average with RISP? I can think of several reasons:

* the pitcher cannot use a wind-up
* the pitcher is unlikely to pitch a breaking ball (curve, splitter or change-up)
* the hitter is more likely to swing just to get a hit, rather than swing for the fences.

I ran the F Test to compare variances of the batting averages when there is no runner on base and when there is runner in scoring position:

F test to compare two variances

data: data\_NO$BA and data\_SP$BA

F = 0.51772, num df = 7899, denom df = 7642, p-value < 2.2e-16

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.4951969 0.5412663

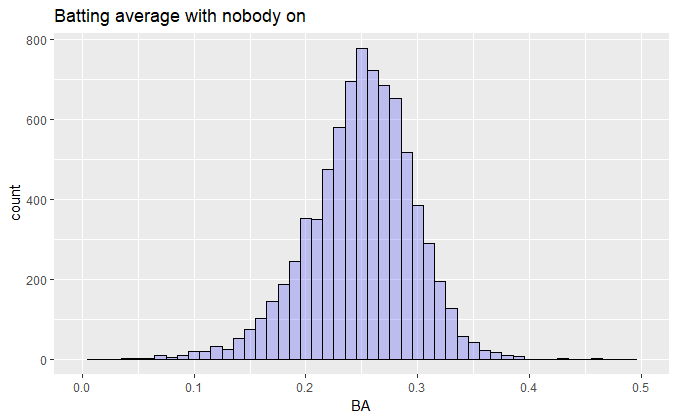
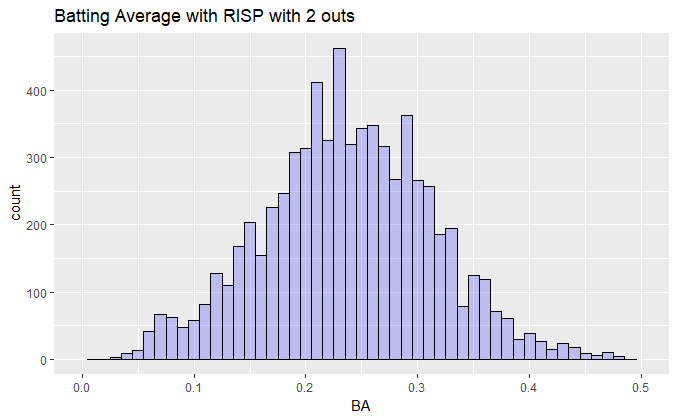
sample estimates:

ratio of variances

0.5177238

The F test indicates that there is enough evidence to reject the null hypothesis that the two sub-group variancess are equal at the 0.01 significance level.

Among the “clutch” situations, the most “stressful” or “pressure” situation is Runner(s) in scoring position with two outs where “clutch hitting” would be required the most. I plotted the distribution of batting averages (number of hits divided by at bats) for these two situations:

This time, the mean batting average when there is runner(s) in scoring position (.238) is statistically significant lower than the batting average when there is no runner on base (.250), p-value<0.01. Again, there is a wider spread of performance when there is Runner(s) in scoring position with two outs. I ran the F Test to compare variances of the batting averages between the two situations above:

F test to compare two variances

data: data\_NO$BA and data\_SP2O$BA

F = 0.34388, num df = 7899, denom df = 6988, p-value < 2.2e-16

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.3285690 0.3598885

sample estimates:

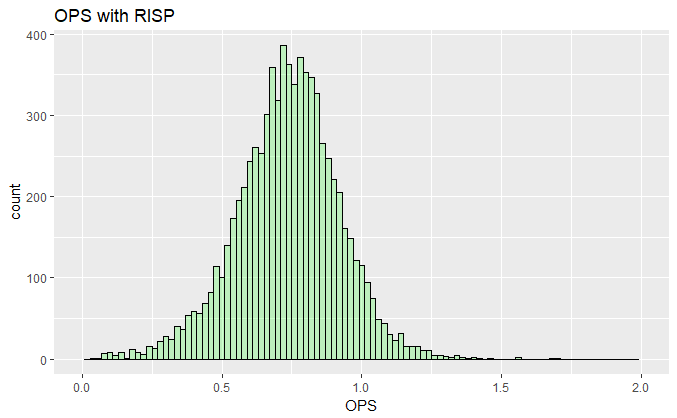
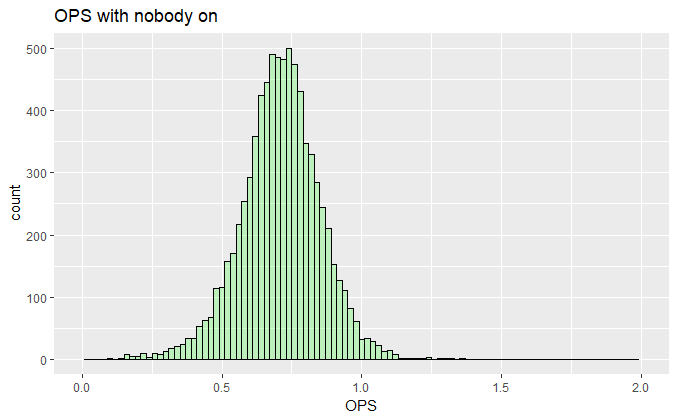
ratio of variances

0.3438834

The F test indicates that there is enough evidence to reject the null hypothesis that the two sub-group variancess are equal at the 0.01 significance level. The wider spread of variances supports the hypothesis that there are indeed some players who play exceptionally well (above average) in clutch situations while some players completely crumble (below average).

However, batting average is a very simplistic way to measure a batter’s ability to hit. Two obvious problems of using batting average arise: 1) every hit has the same weight (a home run is not treated differently from a single) and 2) the patience of drawing a walk in a clutch situation is not factored in the analysis. Therefore, One-Base Plus Slugging (OPS) is a superior statistic to evaluate a hitter's skill, relative to others. OPS is the sum of a player's On-Base Percentage (OBP) and their Slugging Percentage (SLG%). This statistic allows me to examine the two aforementioned components of a hitter: how frequently they get on base (by getting a hit or a walk - and therefore do not make outs), and what type of hitter they are, based on their ability to hit for power.

I plotted the OPS distribution between when there is no runner on base, when there is runner(s) in scoring position and when there is runner(s) in scoring position with two outs. The means are .709 for non-clutch situations and .736 for clutch situations.



There is once again a wider spread for OPS in clutch situations. The F Test confirmed the statistically significant difference between the two sub-groups OPS variances (p-value<0.01):

F test to compare two variances

data: data\_NO$OPS and data\_SP$OPS

F = 0.58857, num df = 7899, denom df = 7642, p-value < 2.2e-16

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.5629631 0.6153369

sample estimates:

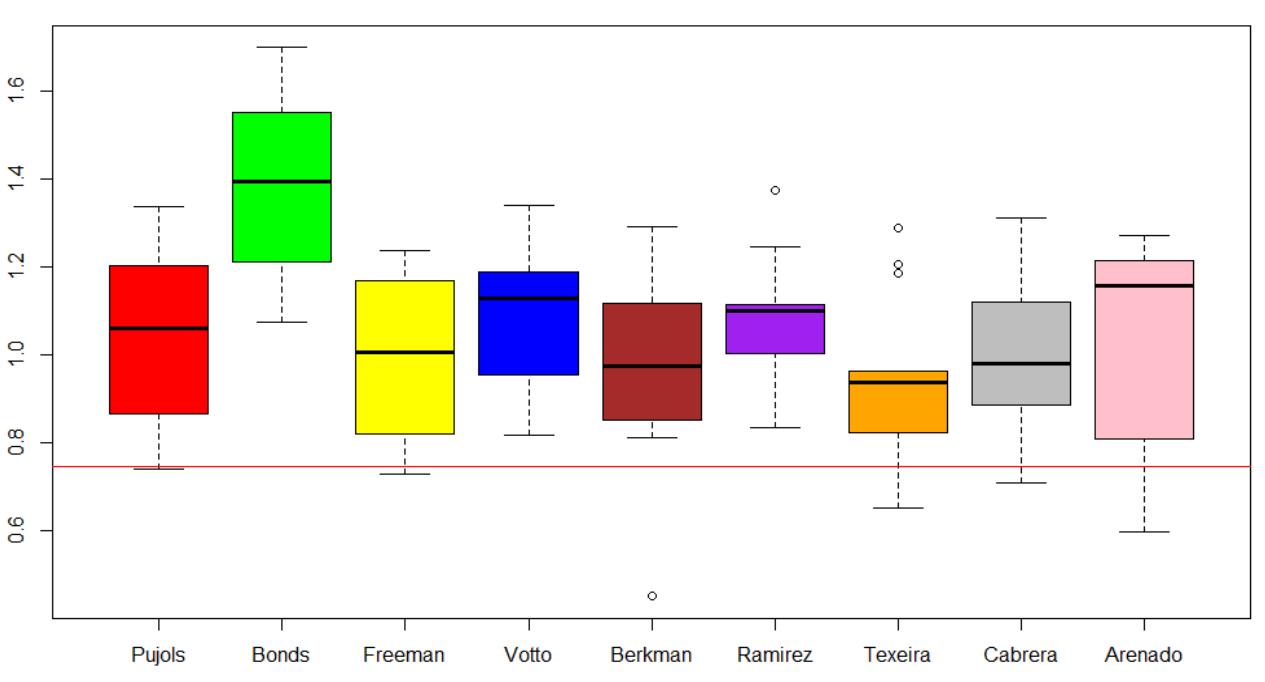
ratio of variances

0.5885726

On an individual level, I ran into another problem with the dataset. There are a number of players who got traded to a different team in the middle of a season and thus their OPS are recorded separately. I developed a function to detect these duplicates and calculate the weight average OPS with the number of at bats as the weight. For example, in 2008 Mark Teixera was having an OPS of 0.832 (no runner on base) with the Atlanta Braves (182 at-bats). Later that year, he was traded to the Los Angeles Angles and had an OPS of 1.243 there (no runner on base, 102 at-bats). As a result, his weighted average OPS when there is no runner on base in 2008 would be 0.980.

Checking the data, I noticed that Alex Gonzalez in 2003 and Luis Gonzalex in 2005 had more than 800 at-bats. This is impossible because Jimmy Rollins holds the record for the number of at-bats in one season (2007) with 716 at-bats. It turned out that there were two players with the name Alex Gonzalez and two players with the name Luis Gonzalez so my function mistakenly merged them into one player. Subsetting these two special cases out from the data before running my algorithm solved this issue.

First, I measured a batter’s “clutchness” strictly by their OPS when there is runner in scoring position. I sorted the top 50 single season performance with the highest OPS in clutch situation. However, some of the batters in this group have less than 150 at-bats and it is highly likely that they benefit from over-inflation due to a small sample size. Hence, I sorted the top 50 single season OPS with at least 158 at-bats (25th percentile). There are 9 players that appear multiple times in this table: Albert Pujols, Barry Bonds, Freddie Freeman, Joey Votto, Lance Berkman, Manny Ramirez, Mark Texeira, Miguel Cabrera and Nolan Arenado. I called these players the Clutch 9: a group of best “clutch” hitters from 2002 to 2017. I plotted a boxplot diagram of the Clutch 9’s career clutch OPS (when there is runner in scoring position). The red line is the league average clutch OPS (.744). None of the lower quartile is lower than the league’s average. In fact, four of the Clutch 9 (Bonds, Votto, Berkman, Ramirez) has a career low clutch OPS higher than the league average.



I suspect Mark Texeira, who has three upper outliers, to not have a statistically significant difference from the league average. However, the one sample t-test showed a statistically significant difference between Mark Teixera clutch OPS and the league average (p-value<0.01):

One Sample t-test

data: Mark\_Texeira

t = 3.581, df = 12, p-value = 0.001888

alternative hypothesis: true mean is greater than 0.7439446

95 percent confidence interval:

0.8406117 Inf

sample estimates:

mean of x

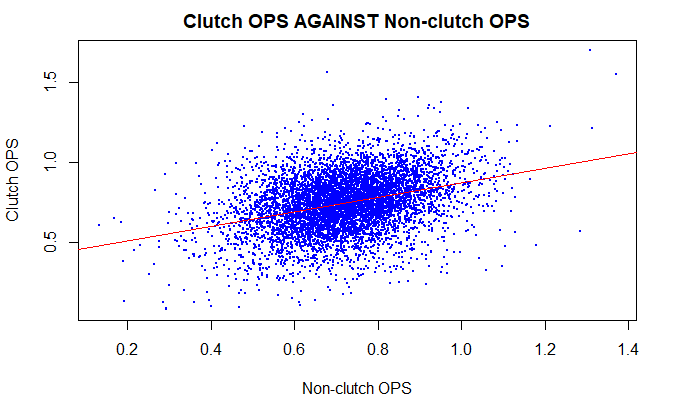
0.9363964

The boxplots showed that clutch players can reproduce their results in a consistent manner and not due to an aberration. Therefore, “clutch hitting,” as an ability to score runners who are in scoring position, can be statistically proven to exist.

Another way to measure a batter’s “clutchness” is to use a more stringent definition: the OPS difference between clutch situations (runner in scoring position) and non-clutch scenarios (no runner on base).

I also sorted the top 50 single season performance with the highest OPS difference. The minimum 158 at-bats requirement is also implemented. If I used the OPS difference as the benchmark, there are four batters who appear multiple times in the top 50: Freddie Freeman, Manny Ramirez, Mike Sweeney and Mark Teixera.

Moreover, I plotted the correlation between Clutch OPS and Non-clutch OPS:



The red line is the best-fit line of the linear regression model. Players who are above the line would be considered as a clutch hitter (high clutch OPS compared to non-clutch OPS). On the other hand, players who are below the line would be considered as a non-clutch hitter.

2) Multiple regression model of clutch hitting

Using clutch OPS (when there is runner in scoring position), I aimed to build a statistical model that could predict a batter’s “clutchness.” Since clutch OPS follows is a continuous variable, I determined from Figure 2 and the Normal Q-Q plot that clutch OPS followed a normal distribution:

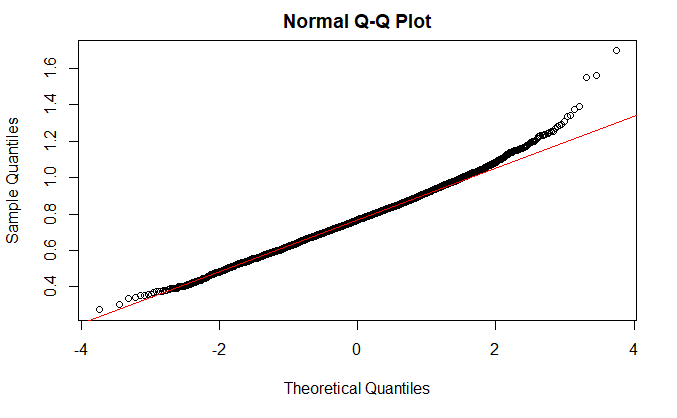
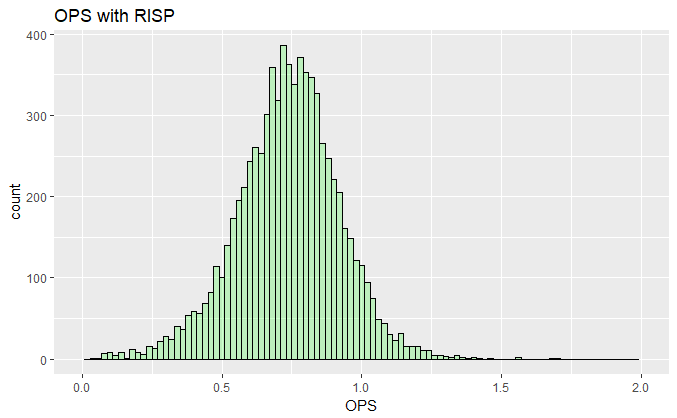


Figure 2 Clutch OPS distribution

The clutch OPS is the response variable. The predictor variables analyzed in this research were as follows:

1. HR (Home Runs): Number of home runs.
2. R (Runs Scored): Number of runs scored.
3. X2B (Doubles): Number of doubles (two-base hit).
4. X3B (Triples): Number of triples (three-base hit).
5. RBI (Runs batted in): Numbers of runs that the batter drives in with hits or walks.
6. TB (Total bases): Number of bases a player has gained with hits or walks.
7. BB (Base on ball): Number of walks.
8. BA (Batting Average): Rate of hits per at bat.
9. OBP (On Base Percentage): Rate at which the batter reaches base.
10. SLG (Slugging Percentage): Average number of total bases per at bat.
11. SO (Strike outs): Number of times a batter is struck out
12. H (Hits): Number of hits

A preliminary investigation into the univariate linear regression models between the 12 variables and clutch OPS revealed a number of interesting correlations:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Estimate** | **P-value** |
| **R** | 0.0026392 | 0e+00 |
| **X2B** | 0.0061344 | 0e+00 |
| **X3B** | 0.0045460 | 2e-07 |
| **HR** | 0.0077167 | 0e+00 |
| **RBI** | 0.0033154 | 0e+00 |
| **TB** | 0.0010002 | 0e+00 |
| **BB** | 0.0031183 | 0e+00 |
| **SO** | 0.0009515 | 0e+00 |
| **BA** | 0.0021548 | 0e+00 |
| **OBP** | 0.0021572 | 0e+00 |
| **SLG** | 0.0012845 | 0e+00 |
| **H** | 0.0013984 | 0e+00 |

Figure 3 List of explanatory variables for clutch OPS

The highlighted statistics were in percentage in the raw data set. I converted the estimates of these statistics for easier interpretation. For example, if a batter increases his batting average by one point (from 0.291 to 0.292), their clutch OPS would increase by 0.0021548 on average (p<0.01).

A surprising observation is that the more a batter is struck out, the higher his clutch OPS would be (on average an 0.00095 increase in clutch OPS for every strike out, p-value<0.01). The popular belief is that strike out is the worst outcome of an at bat. If the goal of a batter is to avoid getting out, strikeout is in fact the third-worst way to end an at-bat (behind double-play or the ultra-rare triple play). Is it possible that getting struck out makes a batter a better clutch hitter? Of course not. My baseball knowledge told me that was logically impossible and it was confirmed by the univariate model.

Correlation does not imply causation. It could be that more strike outs leads to higher cluth OPS, or it could be that there was some third factor that was positively correlated with clutch OPS, but negatively correlated with strike outs. Knowing something about baseball let me argue for which conclusion makes more sense. Hence, I suspected there was a confounder between clutch OPS and Strike outs.

A more accurate way to put the regression result is:

"One extra strike out is associated with higher clutch OPS. That's either because of the strike out itself, or because of something else about players who get struck out a lot, something that was not controlled for in the regression."

One possible explanation for this was Home Run hitters would tend to strike out more. From Figure 3, home run increases clutch OPS by the highest magnitude (0.0077, p-value<0.01). I ran a linear regression to confirm that Home Run hitters would strike out more:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.647433 0.276986 -5.948 2.91e-09 \*\*\*

SO 0.188650 0.003112 60.624 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.577 on 4890 degrees of freedom

(457 observations deleted due to missingness)

Multiple R-squared: 0.4291, Adjusted R-squared: 0.429

F-statistic: 3675 on 1 and 4890 DF, p-value: < 2.2e-16

There is a statistically significant correlation between getting more strike outs and hitting more home runs. In fact, Babe Ruth held the all-time record for career strikeouts for 35 years. Then Mickey Mantle passed him and held the record for 14 years. Willie Stargell passed him and held the record for 4 years. Reggie Jackson then stole the record and has held it for 34 years. All four of those players are inner-circle Hall of Famers.

In practice, using a linear regression with 35 covariates is cumbersome and inconvenient. Hence, I implemented the stepwise regression to select the best model to predict wins. I selected the Akaike Information Criteria (AIC) as the score for candidate models. However, there would certainly be information lost due to using a candidate model to represent the "true" model. As a result, I aimed to select, from among the candidate models, the model that minimizes the information loss. I implemented backward stepwise elimination, which involved starting with all candidate variables, testing the deletion of each variable using AIC, deleting the variable (if any) whose loss gave the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit. Below are the results of my backward stepwise AIC elimination of variables:

glm(formula = OPS\_SP ~ RBI + SO + OBP + SLG, data = data\_overall\_clutch\_158)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.36645 -0.07041 -0.00620 0.06455 0.78468

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.053e-01 1.426e-02 14.40 <2e-16 \*\*\*

RBI 2.190e-03 9.610e-05 22.79 <2e-16 \*\*\*

SO -6.909e-04 5.598e-05 -12.34 <2e-16 \*\*\*

OBP 8.707e-01 5.501e-02 15.83 <2e-16 \*\*\*

SLG 5.043e-01 3.717e-02 13.57 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.01164346)

Null deviance: 111.486 on 4891 degrees of freedom

Residual deviance: 56.902 on 4887 degrees of freedom

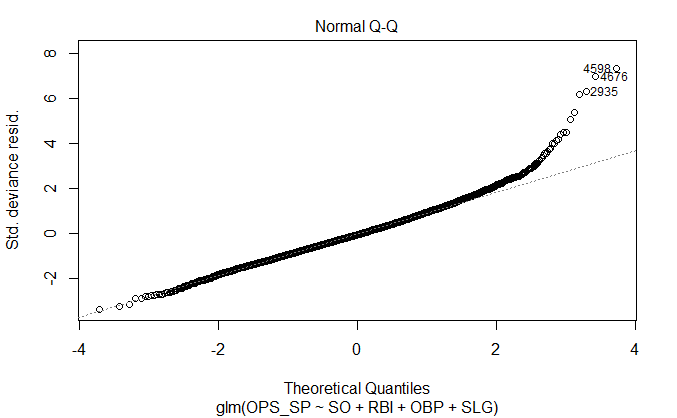
(457 observations deleted due to missingness)

AIC: -7894.2

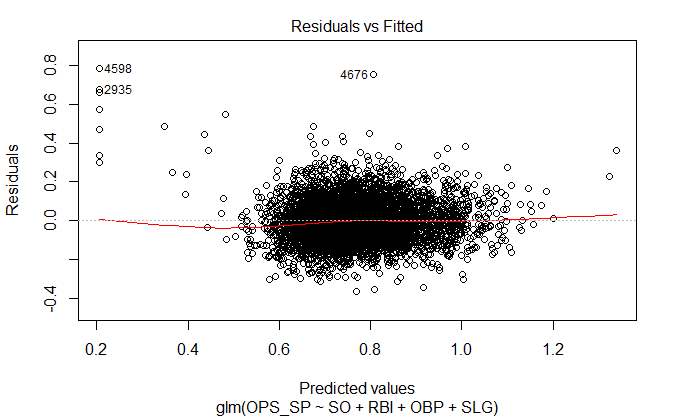
Number of Fisher Scoring iterations: 2

The final model is clutch OPS ~ 0.00219\*RBI - 0.0006909\*SO + 0.8707\*OBP + 0.5043\*SLG.

The Normal QQ plot indicated that the residuals of the simplified model are normally distributed:



The Residuals vs Fitted plot showed no pattern to the residuals. They appeared to be flat and not have a difference in width of the range of values. This would suggest homoscedasticity of residuals from my simplified linear regression model.



I performed a Hosmer-Lemeshow test for Goodness-of-fit of the simplified model. The hypotheses for the test are stated as follows:

Hosmer and Lemeshow goodness of fit (GOF) test

data: data\_overall\_clutch\_158$OPS\_SP, fitted(full\_model)

X-squared = 1.0973, df = 8, p-value = 0.9976

With a very large p-value (p-value=1) I failed to reject the null hypothesis. This means that my predicted and actual events are very well distributed across our 10 groups. Therefore, this simplified model could be concluded as a very good fit to the data.

This simplified model might face the problem of multicollinearity between Runs Batted In (RBI) and Slugging Percentage (SLG). Multicollinearity arises when there are high correlations among predictor variables, leading to unreliable and unstable estimates of regression coefficients. I used the variance inflation factor as a diagnostic for multicollinearity. The VIF is calculated for each predictor by doing a linear regression of that predictor on all the other predictors, and then obtaining the R-squared from that regression. The VIF is computed as 1/(1-R-squared). Authorities differ on how high the VIF has to be to constitute a problem. However, a VIF lower than 2.5 generally means that multicollinearity can be safely ignored.

Call:

lm(formula = RBI ~ SLG, data = data\_overall\_clutch\_158)

Residuals:

Min 1Q Median 3Q Max

-72.870 -12.346 0.025 12.762 65.111

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -56.771 1.515 -37.47 <2e-16 \*\*\*

SLG 265.281 3.520 75.36 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.43 on 4890 degrees of freedom

(457 observations deleted due to missingness)

Multiple R-squared: 0.5373, Adjusted R-squared: 0.5372

F-statistic: 5679 on 1 and 4890 DF, p-value: < 2.2e-16

The VIF between Slugging Percentage (SLG) and Runs Batted In (RBI) is 2.16.

Based on the simplified model (clutch OPS ~ 0.00219\*RBI - 0.0006909\*SO + 0.8707\*OBP + 0.5043\*SLG), a baseball executive can easily use the statistics Runs Batted In (RBIs), Strike outs(SO), On base percentage (OBP) and Slugging percentage (SLG) to select which player to sign due to their straightforward nature. In contrast, using traditional measurements such as Home Runs (R) required further analysis because of the superficial implications they offered. In this regard, modern sabermetrics are superior to traditional statistics.

An inherent weakness of my model lied in the assumptions of logistic regression. The problem with the assumptions behind ordinary logistic regression is that I treated each team-season observation as independent, regardless of whether the team had the same player (roster changes, players’ aging, etc.) or the same season in which observations occurred. This can be dangerous because some teams traded actively during mid-season (July) or called up prospects, so team roster changes by month observations and the underlying performance function was not constant throughout the season. Additionally, a player’s capability changes from season to season, adding another layer of complexity as the 2015 Yankees would not play like the 2016 Yankees even with no roster change due to decaying performance of older players. A way to fix this is to use survival analysis instead. However, the limited data (and time) on my hands did not allow such endeavor.

3) How clutch hitting affects a baseball team performance?

I investigated the relationship between a team’s clutch hitting (OPS with runner in scoring position) and the number of their wins. I collected the data from 18 seasons (2000 to 2017).

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.617 6.937 1.242 0.215

OPS 95.321 9.112 10.462 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 108.0059)

Null deviance: 71224 on 551 degrees of freedom

Residual deviance: 59403 on 550 degrees of freedom

AIC: 4155.1

Number of Fisher Scoring iterations: 2

There is a statistically significant (p-value<0.01) that every increase in 100 points of clutch OPS (.100) would on average increase a team’s performance in a 162-game season by 9.53 wins. A 9.53 increase in wins would be a 91-71 record in a 162-game baseball season.

Using a binomial simulation (10,000 times) estimated that an increase in 100 points of clutch OPS (.100) will increase a team’s chance of winning a baseball game by 6.1% (from 0.5 to 0.561).

