T. Dileepa Ariyawansa

13 December 2019

My project is going to be about using certain statistics to find which players are due to breakout and which players are due for regression. The statistics I will be using will be wOBA (Weighted On Base Percentage), xwOBA (Expected Weighted On Base Percentage) and OPS (On Base Plus Slugging). Using these statistics I will find the best 9 batters for the lowest cost for the upcoming 2020 season using data from 2015 - 2019 to prove my hypothesis. This topic interests me as I've always loved baseball since I was 5 years old and always was fascinated by how deep statistically baseball is. Baseball was one of huge reasons why I became a Statistics and Economics major and Math minor in college as it made me want to explore the world of Statistics and Economics more deeply.

The data I will be using wOBA (Weighted On Base Percentage), xwOBA (Expected Weighted On Base Percentage) and OPS (On Base Plus Slugging) I will find which players are due for a breakout and which players are due for a regression. I will find the data using multiple websites like baseballsavant.mlb.com and fangraphs.com to find the wOBA, xwOBA and OPS. The econometrics model in this project is finding the difference between xwOBA and wOBA to possible determine a players future production. Subtracting xwOBA from wOBA will then get rid of Omitted Variable Bias as it gets rid of luck. The parameter of interest is the difference between xwOBA and wOBA (diff = xwOBA - wOBA).

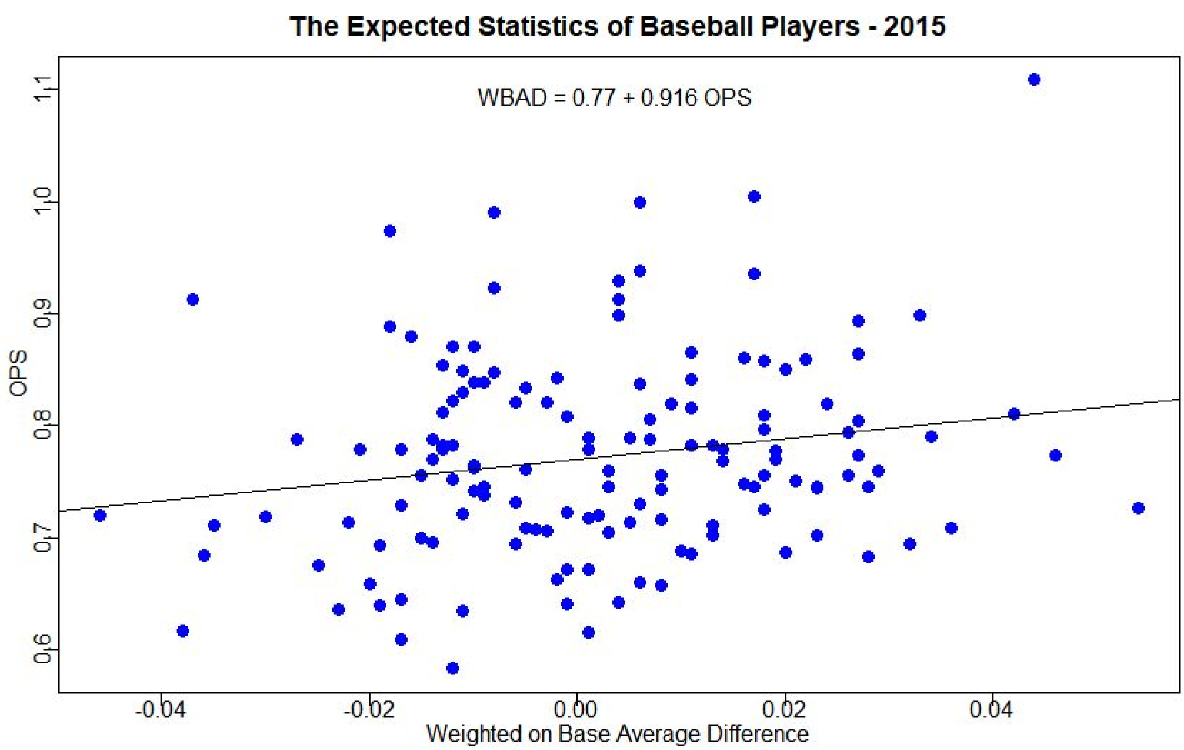
I will be using OLS to estimate the model by using a players diff (xwOBA - wOBA) for the independent variable and the players current year OPS for the dependent variable. I will also conduct a normal z test on the model to see whether the findings support my hypothesis or not.

These are the results of the findings:

```
2015:
Residuals:
     Min       1Q   Median       3Q      Max
-0.17603 -0.05956 -0.01080  0.05582  0.29867

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.770026   0.007436 103.557   <2e-16 ***
wobadif     0.915920   0.402101   2.278   0.0243 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08801 on 140 degrees of freedom
Multiple R-squared:  0.03574,	Adjusted R-squared:  0.02885
F-statistic: 5.189 on 1 and 140 DF,  p-value: 0.02425
```



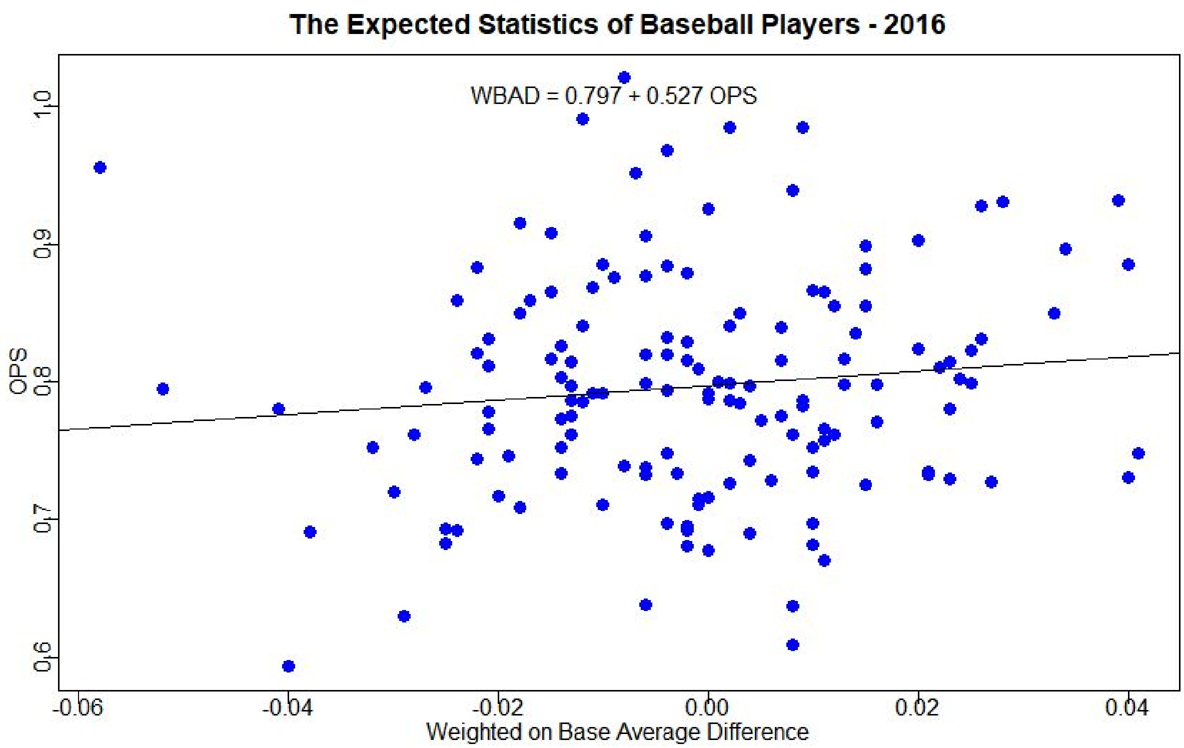
2016:
Residuals:

```
      Min        1Q    Median        3Q       Max
-0.191990 -0.059822 -0.002903  0.048486  0.228442
```

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.796774   0.006748 118.078   <2e-16 ***
wobadif     0.527017   0.370818   1.421    0.157
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.0814 on 144 degrees of freedom
Multiple R-squared: 0.01383,	Adjusted R-squared: 0.006985
F-statistic: 2.02 on 1 and 144 DF, p-value: 0.1574



2017
Residuals:

```
      Min        1Q    Median        3Q       Max
-0.201065 -0.063917 -0.005373  0.055106  0.262218
```

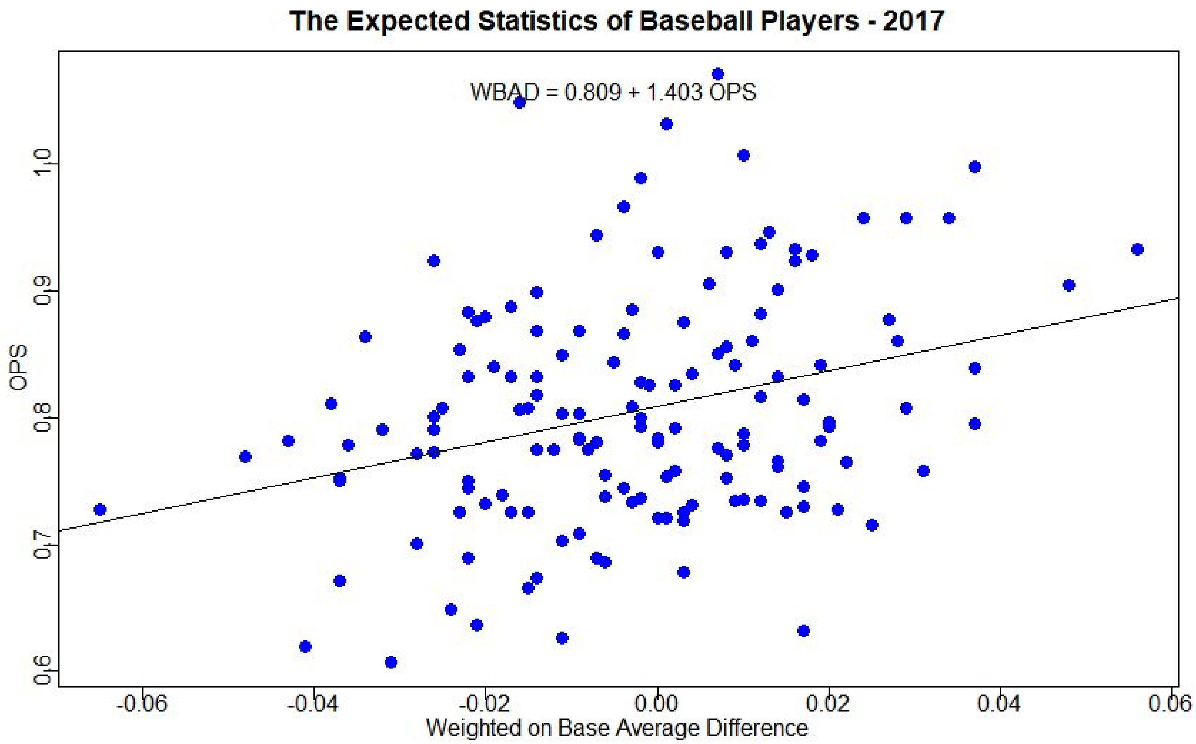
Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.809222   0.007298 110.879  < 2e-16 ***
```

wobadif   1.402507   0.358960   3.907 0.000144 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0868 on 142 degrees of freedom
Multiple R-squared: 0.09707,	Adjusted R-squared: 0.09071
F-statistic: 15.27 on 1 and 142 DF, p-value: 0.000144

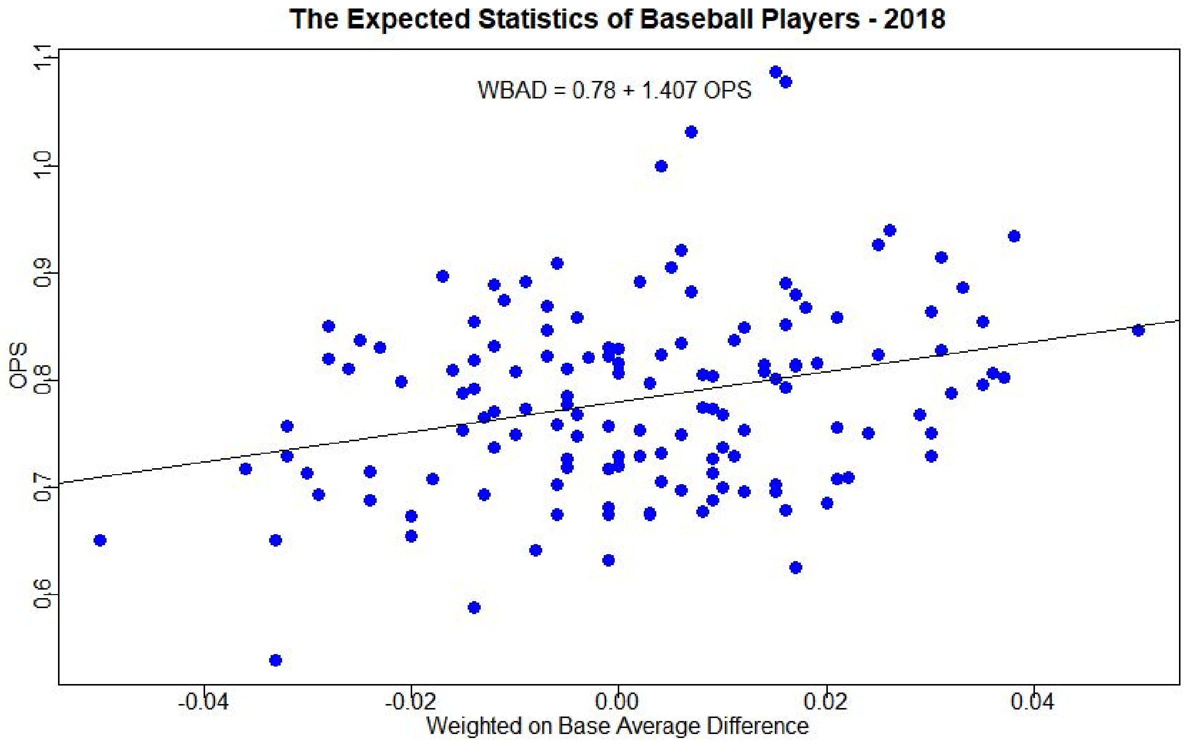


2018
Residuals:
     Min   1Q   Median   3Q   Max
-0.194999 -0.060436 -0.004795 0.051079 0.285456

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.780436   0.007281 107.194 < 2e-16 ***
wobadif   1.407178   0.397018   3.544 0.000537 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08578 on 139 degrees of freedom
Multiple R-squared: 0.08289,	Adjusted R-squared: 0.07629
F-statistic: 12.56 on 1 and 139 DF, p-value: 0.0005366



2019
Residuals:

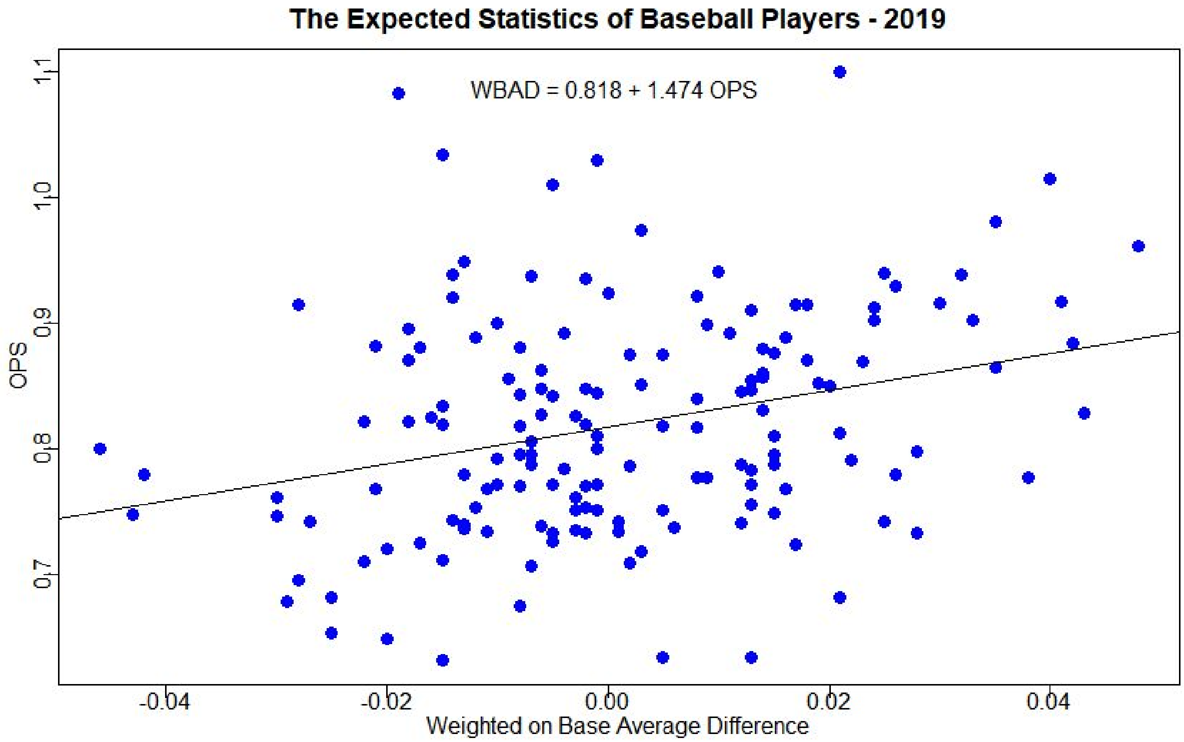
Min	1Q	Median	3Q	Max
-0.20271	-0.06221	-0.00653	0.05064	0.29347

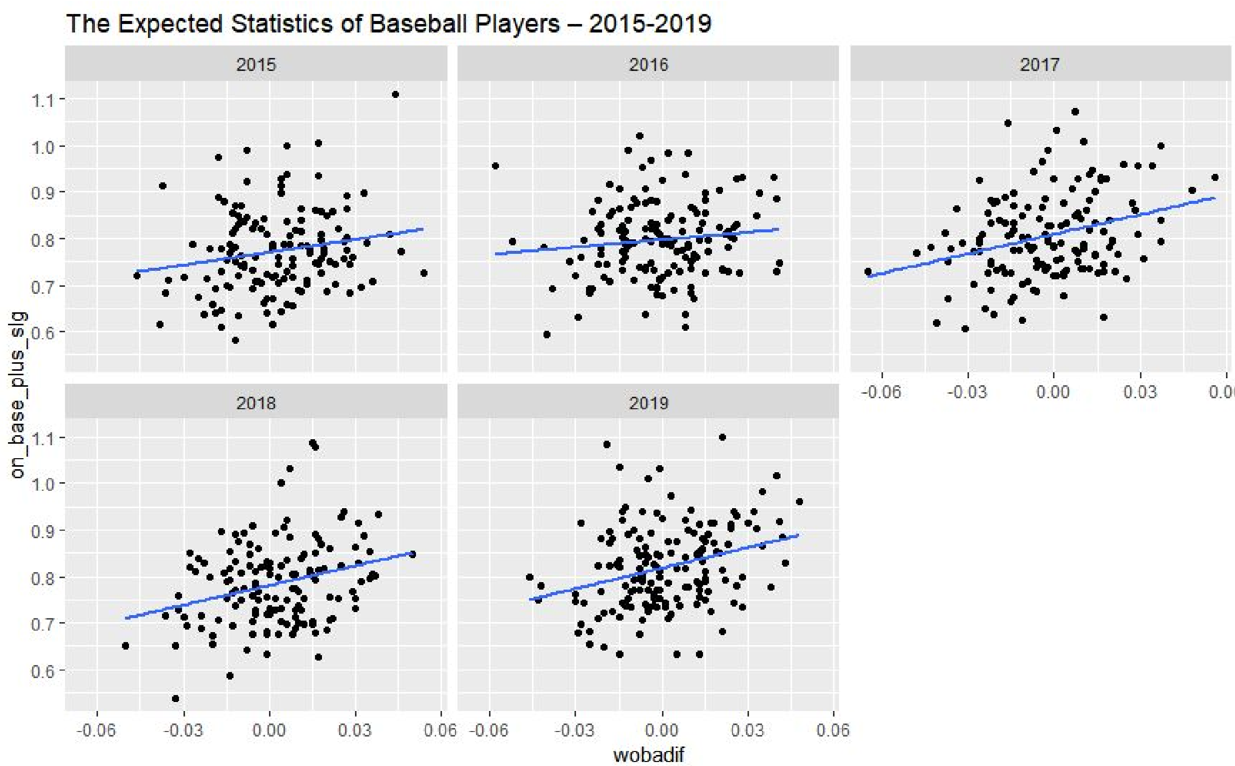
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.8175	0.0069	118.491	< 2e-16	***
wobadif	1.4744	0.3792	3.889	0.00015	***

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

Residual standard error: 0.08593 on 154 degrees of freedom
Multiple R-squared: 0.08941, Adjusted R-squared: 0.0835
F-statistic: 15.12 on 1 and 154 DF, p-value: 0.0001495





The results of the experiment show that there does not seem to be a clear cut relationship between the difference between xwOBA and wOBA and a players given OPS. Since the P value does not equal zero we cannot reject the null hypothesis. Though there seems to be a noted relationship in the data with a majority of players previous years xwOBA to a players current OPS. This model is externally valid as we can use this model across multiple leagues other than the MLB (MILB, KBO, NPB etc) while this model may suffer from internal invalidity as it did not take into account a players age and injury status (some players play through injury therefore their stats may be hindered).

Based on the data type I would consider using Panel Data Regression as my data spans multiple years and uses multiple variables to help calculate my model. There is not a good instrument variable I can think of that can be implemented in the IV regression techniques.