Player Evaluation using wRC+

And writing a new model to compare MLB batters

Taylor Stacey
Dr. Rajarshi Dey

August 10, 2021

Emporia University

Table of contents

Research Question

Research Question

- The objective is to find a model using Major League Baseball career data for players dating back to 1871 to explain variation in career wRC+

Definitions

Statcast

- It is a tracking technology system that provides collection and analysis for large amounts of baseball data.
- Each MLB stadium has 13 Hawk-Eye camera systems, five for pitch tracking and seven for tracking players and batted balls.
- The data obtained is useful for front offices, broadcasters, and baseball fans to find a new level of understanding about the skills of players on the field.
- Some measurements observed by statcast include: pitcher spin rate, direction, and movement, batter exit velocity, launch angle, barrel percentage, and batted ball distance, and fielding arm strength, catch probability, and catcher pop time.

Statcast

Figure 1 below illustrates some data provided by Statcast which leads to new measures.

layer	Year	wOBA	xwOBA	wOBA - xwOBA	Avg EV (MPH)	Avg LA (°)	Sweet Spot %	Barrel%	Solid Contact %
Perez, Salvador	2021	.356	.367	-0.011	92.8	14.8	34.9	14.4	9.6
Baez, Javier	2021	.310	.297	.013	90.1	12.1	29.3	13.6	8.2
Iglesias, Jose	2021	.290	.291	-0.001	86.2	9.3	32.3	4.2	3.1

Figure 1: Statcast Example

Some Statcast Measurements

- Spin Rate
  - How much spin, in revolutions per minute, a pitch was thrown with.
- Launch Angle
  - How high, in degrees, a ball was hit by a batter.
- Exit Velocity
  - How fast, in miles per hour, a ball was hit by a batter.

Some Statcast Metrics

- Barrels
  - A batted ball with the perfect combination of exit velocity and launch angle, or the most high-value batted balls. (A barrel has a minimum Expected Batting Average of .500.)
- Catch Probability
  - The likelihood, in percent, that an outfielder will be able to make a catch on an individual batted ball. Catch Probability accounts for distance needed, time available, direction, and proximity to the wall, compared to how often the same opportunity is caught by Major League outfielders.
- Sprint Speed
  - A measurement of a player's top running speed, expressed in ”feet per second in a player's fastest one-second window.”

Player Evaluation Metrics

FANGRAPHS

Weighted Runs Created Plus (wRC+) is a rate statistic which attempts to credit a hitter for the value of each outcome (single, double, etc) rather than treating all hits or times on base equally, while also controlling for park effects and the current run environment. wRC+ is scaled so that league average is 100 each year and every point above or below 100 is equal to one percentage point better or worse than league average. This makes wRC+ a better representation of offensive value than batting average, RBI, OPS, or wOBA.

$$wRC+ \; = \frac{(wRAA/PA + LgR/PA) + (LgR/PA - (Park\ Factor \times LgR/PA))}{(AL\ or\ NL\ wRC/PA\ excluding\ pitchers)} \times 100$$

Figure 2: WRC+ explained by Fangraphs

wRC+ Formula Explained

$$\text{wRC+} = \frac{(\frac{wRAA}{PA} + \frac{LgR}{PA}) + (\frac{LgR}{PA} - (\text{Park Factor} * \frac{LgR}{PA}))}{(\text{AL or NL} \frac{wRC}{PA} \text{excluding pitchers})} * 100$$

- wRAA - weighted runs above average, measures the number of offensive runs a player contributes to their team compared to the average player.
- PA - plate appearances.
- LgR - league runs.
- Park Factor - ballpark factors, this measures how the rate of difficulty at ballparks varies depending on the environment and landscape of the individual ballpark.
- wRC - weighted runs created, a measure to quantify a player's complete offensive value in runs scored.

A Closer Look at Park Factor

Season	Team	Basic (5yr)
2020	Angels	99
2020	Orioles	100
2020	Red Sox	104
2020	White Sox	99
2020	Indians	103
2020	Tigers	102
2020	Royals	102
2020	Twins	101
2020	Yankees	100
2020	Athletics	96
2020	Mariners	96
2020	Rays	96
2020	Rangers	100
2020	Blue Jays	103
2020	Diamondbacks	101
2020	Braves	101
2020	Cubs	99
2020	Reds	102
2020	Rockies	114
2020	Marlins	95
2020	Astros	96
2020	Dodgers	95
2020	Brewers	100
2020	Nationals	102
2020	Mets	95
2020	Phillies	100
2020	Pirates	99
2020	Cardinals	96
2020	Padres	97
2020	Giants	97

Figure 3: WRC+ Comparison by Fangraphs (Statistics for 2021 season as of 6/28)

wRC+ Scale

Ratings	wRC	wRC+
Excellent	105	160
Great	90	140
Above Average	75	115
Average	65	100
Below Average	60	80
Poor	50	75
Awful	40	60

Figure 4: WRC+ Scale by Fangraphs

wRC+ Player Comparison

#	Name	Team	wRC+
1	Jake Cronenworth	SDP	131
2	Kyle Seager	SEA	93
3	J.P. Crawford	SEA	110
4	Matt Chapman	OAK	105
5	Isiah Kiner-Falefa	TEX	91
6	Cedric Mullins II	BAL	151
7	Nate Lowe	TEX	116
8	Dansby Swanson	ATL	94
9	Elvis Andrus	OAK	59
10	Vladimir Guerrero Jr.	TOR	200

Figure 5: WRC+ Comparison by Fangraphs (Statistics for 2021 season as of 6/28)

wRC+ Explained by Statcast Variables

$$\text{wRC+} = -106.881 + 2.954 * \text{Speed Score} + 183.283 * \text{Line Drive\%} + 1.805 * \text{Exit Velocity} + 448.981 * \text{Barrel\%} + 166.706 * \text{Walk\%} - 224.558 * \text{Strikeout\%}$$

- Data analysis done by Ryan Kupeic in his paper ”Can Statcast variables explain the variation in weighted runs created plus?”
- Data set included 406 players from the 2019 MLB Season. (Last full season)
- Each regressor is significant in the model at the 0.05 level.
- Adjusted R-Squared is 0.679 and Residual Standard Error is 15.63.

wRC+ Explained by Statcast Variables

wRC+$=-106.881+2.954*$Speed Score$+183.283*$Line Drive%$+1.805*$Exit Velocity$+448.981*$Barrel%$+166.706*$Walk%$-224.558*$Strikeout%

- The original model by Kupeic contained 11 variables with two variables not significant (opposite field percentage and Speed Score) and two variables with high Variance Inflation Factors (Flyball percentage and Launch angle).
- To find best model, stepwise procedures are used, which includes forward selection, backward elimination, and stepwise regression.
- The goal is to find the highest adjusted R-Squared and lowest mallow's CP and AIC with no multicollinearity and fewest variables possible.
- Finally, before we accept our best model, we need to check our model assumptions by looking at residual plots and QQ-plot.

wOBA

FANGRAPHS

Weighted On-Base Average (wOBA) is a rate statistic which attempts to credit a hitter for the value of each outcome (single, double, etc) rather than treating all hits or times on base equally. wOBA is on the same scale as On-Base Percentage (OBP) and is a better representation of offensive value than batting average, RBI, or OPS. The weights change slightly with the run environment, but the general formula is:

$$wOBA = \frac{.69 \times uBB + .72 \times HBP + .89 \times 1B + 1.27 \times 2B + 1.62 \times 3B + 2.10 \times HR}{AB + BB - IBB + SF + HBP}$$

Figure 6: wOBA explained by Fangraphs

A New wRC+ Model Using Using 1871 to 2021 Career Data

Data Set Used For Model

1 2 3 4 5 6 7 8 9 10 ... Page size: 30 — 4099 items in 137 pages

#	Name	Team	G	PA	HR	R	RBI	SB	BB%	K%	ISO	BABIP	AVG	OBP	SLG	wOBA	xwOBA	wRC+	BsR	Off	Def	WAR
1	Babe Ruth	- - -	2503	10616	714	2174	2217	123	19.4%	12.5%	.348	.340	.342	.474	.690	.513		197	-23.4	1347.3	-18.6	168.4
2	Barry Bonds	- - -	2986	12606	762	2227	1996	514	20.3%	12.2%	.309	.285	.298	.444	.607	.435		173	30.4	1173.8	67.6	164.4
3	Willie Mays	- - -	2992	12493	660	2062	1903	338	11.7%	12.2%	.256	.299	.302	.384	.557	.409		154	32.9	837.5	170.1	149.9
4	Ty Cobb	- - -	3035	13072	117	2246	1937	892	9.6%	4.1%	.146	.378	.366	.433	.512	.445		165	60.6	1036.0	-90.0	149.3
5	Honus Wagner	- - -	2792	11739	101	1736	1732	722	8.2%	7.6%	.139	.318	.327	.391	.466	.408		147	56.9	704.7	184.4	138.1
6	Hank Aaron	- - -	3298	13940	755	2174	2297	240	10.1%	9.9%	.250	.291	.305	.374	.555	.403		153	24.9	882.0	-61.2	136.3
7	Tris Speaker	- - -	2789	11988	117	1882	1529	432	11.5%	2.3%	.156	.350	.345	.428	.500	.436		157	4.1	815.2	24.4	130.6
8	Ted Williams	BOS	2292	9791	521	1798	1839	24	20.6%	7.2%	.289	.328	.344	.482	.634	.493		188	-1.6	1064.5	-125.1	130.4
9	Rogers Hornsby	- - -	2259	9475	301	1579	1584	135	11.0%	7.2%	.218	.365	.358	.434	.577	.459		173	-1.8	862.1	126.5	130.3
10	Stan Musial	STL	3026	12712	475	1949	1951	78	12.6%	5.5%	.228	.320	.331	.417	.559	.435		158	6.0	901.2	-77.6	126.8
11	Eddie Collins	- - -	2826	12037	47	1821	1300	744	12.5%	3.2%	.096	.343	.333	.424	.429	.409		144	42.3	663.4	68.3	120.5
12	Lou Gehrig	NYY	2164	9660	493	1888	1995	102	15.6%	8.2%	.292	.332	.340	.447	.632	.477		173	-27.2	954.0	-90.7	116.3
13	Alex Rodriguez	- - -	2784	12207	696	2021	2086	329	11.0%	18.7%	.255	.314	.295	.380	.550	.395		141	35.4	665.1	69.0	113.7
14	Mickey Mantle	NYY	2401	9909	536	1677	1509	153	17.5%	17.3%	.259	.318	.298	.421	.557	.428		170	21.8	842.6	-78.1	112.3
15	Mel Ott	NYG	2730	11337	511	1859	1860	89	15.1%	7.9%	.229	.294	.304	.414	.533	.430		156	12.5	810.3	-42.2	110.5
16	Mike Schmidt	PHI	2404	10062	548	1506	1595	174	15.0%	18.7%	.260	.280	.267	.380	.527	.395		147	-0.7	538.3	150.7	106.5
17	Rickey Henderson	- - -	3081	13346	297	2295	1115	1406	16.4%	12.7%	.140	.305	.279	.401	.419	.372		132	144.4	650.9	-56.1	106.3
18	Frank Robinson	- - -	2808	11743	586	1829	1812	204	12.1%	13.0%	.243	.295	.294	.389	.537	.404		153	15.5	731.6	-131.5	104.0
19	Nap Lajoie	- - -	2480	10460	83	1504	1599	380	4.9%	4.0%	.128	.295	.338	.380	.467	.401		144	-3.0	543.8	86.3	102.2
20	Jimmie Foxx	- - -	2317	9670	534	1751	1922	87	15.0%	13.6%	.284	.336	.325	.428	.609	.460		158	-18.6	761.7	-54.2	101.8
21	Joe Morgan	- - -	2649	11329	268	1650	1133	689	16.5%	9.0%	.156	.278	.271	.392	.427	.372		135	79.0	525.5	14.0	98.8
22	Eddie Mathews	- - -	2391	10101	512	1509	1453	68	14.3%	14.7%	.238	.273	.271	.376	.509	.389		143	3.6	532.7	62.8	96.1
23	Carl Yastrzemski	BOS	3308	13991	452	1816	1844	168	13.2%	10.0%	.177	.290	.285	.379	.462	.375		130	-6.7	463.3	-0.5	94.8
24	Cal Ripken	BAL	3001	12883	431	1647	1695	36	8.8%	10.1%	.172	.277	.276	.340	.447	.346		112	-11.0	165.4	310.1	92.5
25	Cap Anson	- - -	2523	11319	97	1996	2076	276	8.7%	2.8%	.112	.339	.333	.393	.445	.393		134	-35.3	567.1	64.2	91.2

Figure 7: Data set used provided by Fangraphs

Description of Predictors In The Model

- Games
    - Number of games played in which the player has appeared.
- Plate Appearances
    - Number of times the player has come to the plate.
- Home Runs
    - Number of home runs.
- Runs
    - Number of runs scored.

Description of Predictors In The Model

- Runs Batted In (RBI)
  - Number of times a run scores as a result of a batter's plate appearance, not counting situations in which an error caused the run to score or the batter hit into a double play.
- Batting Average on Balls In Play (BABIP)
  - The rate at which the batter gets a hit when he puts the ball in play, calculated as (H-HR)/(AB-K-HR+SF).
- On-Base Percentage (OBP)
  - Rate at which the batter reaches base, calculated as (H+BB+HBP)/(AB+BB+HBP+SF).

Description of Predictors In The Model

- Weighted On Base Average (wOBA)
    - Combines all the different aspects of hitting into one metric, weighting each of them in proportion to their actual run value. While batting average, on-base percentage, and slugging percentage fall short in accuracy and scope, wOBA measures and captures offensive value more accurately and comprehensively.
- Wins Above Replacement
    - A comprehensive statistic that estimates the number of wins a player has been worth to his team compared to a freely available player such as a minor league free agent.
- Offensive Runs Above Average (Off)
    - Number of runs above or below average a player has been worth offensively, combining Batting Runs and BsR.

Original Predictors Used For Model

Games, Plate Appearances, Home Runs, Runs, RBI, BABIP, On-base Percentage, wOBA, WAR, Off

$$\text{wRC+} = \beta_0 + \beta_1 \text{G} + \beta_2 \text{PA} + \beta_3 \text{HR} + \beta_4 \text{R} + \beta_5 \text{RBI} + \beta_6 \text{BABIP} + \beta_7 \text{OBP} + \beta_8 \text{wOBA} + \beta_9 \text{WAR} + \beta_{10} \text{Off}$$

- A few variables were removed initially because they were not significant at the 0.05 level -> SB, ISO, AVG, SLG

Use of CrPlots

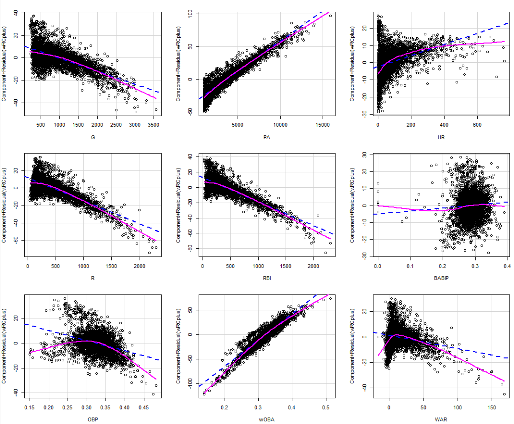

Figure 8: Second order variables for OBP and WAR

$$\text{wRC+} = \beta_0 + \beta_1 \text{G} + \beta_2 \text{PA} + \beta_3 \text{HR} + \beta_4 \text{R} + \beta_5 \text{RBI} + \beta_6 \text{BABIP} + \beta_7 \text{OBP} + \beta_8 \text{wOBA} + \beta_9 \text{WAR} + \beta_{10} \text{Off} + \beta_{11} \text{I}(\text{OBP}^2) + \beta_{12} \text{I}(\text{WAR}^2)$$

Some Information on our Model So Far

```
Call:
lm(formula = wRCplus ~ G + PA + HR + R + RBI + BABIP + OBP +
    wOBA + WAR + Off + I(OBP^2) + I(WAR^2))

Residuals:
    Min      1Q  Median      3Q     Max
-24.881  -4.160  -0.151   3.531  33.462

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.458e+02  4.768e+00 -30.583  < 2e-16 ***
G           -8.014e-03  1.188e-03  -6.745 1.75e-11 ***
PA           6.307e-03  4.545e-04  13.875  < 2e-16 ***
HR           1.872e-02  2.635e-03   7.105 1.42e-12 ***
R           -1.968e-02  1.757e-03 -11.206  < 2e-16 ***
RBI         -2.474e-02  1.538e-03 -16.088  < 2e-16 ***
BABIP        4.266e+00  4.760e+00   0.896     0.37
OBP          6.946e+02  3.521e+01  19.726  < 2e-16 ***
wOBA         4.420e+02  1.088e+01  40.605  < 2e-16 ***
WAR          2.639e-01  2.489e-02  10.599  < 2e-16 ***
Off          9.326e-02  2.736e-03  34.088  < 2e-16 ***
I(OBP^2)    -1.208e+03  5.264e+01 -22.955  < 2e-16 ***
I(WAR^2)    -3.707e-03  1.688e-04 -21.961  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.841 on 3964 degrees of freedom
  (120 observations deleted due to missingness)
Multiple R-squared:  0.9145,    Adjusted R-squared:  0.9142
F-statistic:  3531 on 12 and 3964 DF,  p-value: < 2.2e-16
```

Figure 9: Summary output for Model using R.

- We want to reduce the number of variables.

Stepwise Regression

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	HR	0.8951	0.8948	904.8617	28205.8850	7.5533
2	R	0.9093	0.9091	230.6120	27610.4549	7.0230
3	RBI	0.9136	0.9134	26.3159	27411.3913	6.8536
4	OBP	0.9148	0.9146	-28.0463	27356.6980	6.8072
5	WOBA	NA	NA	NA	NA	NA
6	WAR	NA	NA	NA	NA	NA
7	Off	NA	NA	NA	NA	NA
8	I(OBP^2)	NA	NA	NA	NA	NA
9	I(WAR^2)	NA	NA	NA	NA	NA
10	PA	NA	NA	NA	NA	NA
11	G	NA	NA	NA	NA	NA

Figure 10: ols step forward p using R.

- Variables to keep in Model: HR, R, RBI, OBP.

Stepwise Regression

Selection Summary

Variable	AIC	Sum Sq	RSS	R-Sq	Adj. R-Sq
wOBA	29503.561	1901068.074	321235.038	0.85545	0.85541
BABIP	28633.054	1857253.897	311157.962	0.85650	0.85643
Off	28022.992	1901638.270	266773.589	0.87697	0.87688
I(WAR^2)	27715.410	1921617.099	246794.761	0.88619	0.88607
WAR	27553.715	1931569.038	236842.822	0.89078	0.89064
I(OBP^2)	27499.214	1934910.046	233501.814	0.89232	0.89215
OBP	26978.088	1963690.069	204721.791	0.90559	0.90542
RBI	26860.811	1969738.875	198672.985	0.90838	0.90819
PA	26792.269	1973231.729	195180.131	0.90999	0.90979
R	26675.741	1978962.941	189448.919	0.91263	0.91241
HR	26639.402	1980780.503	187631.357	0.91347	0.91323
G	26596.014	1982909.706	185502.153	0.91445	0.91419

Figure 11: ols step forward aic using R.

- Variables to keep in Model: All.

Stepwise Regression

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	BABIP	0.9148	0.9146	-28.0463	27356.6980	6.8072

Figure 12: ols step backward p using R.

- Variables to remove in Model: BABIP.

Stepwise Regression

```
[1] "No variables have been removed from the model."
```

Figure 13: ols step backward aic using R.

- Variables to remove in Model: None.

Stepwise Regression

Stepwise Summary

Variable	Method	AIC	RSS	Sum Sq	R-Sq	Adj. R-Sq
wOBA	addition	29503.561	321235.038	1901068.074	0.85545	0.85541
BABIP	addition	28633.054	311157.962	1857253.897	0.85650	0.85643
Off	addition	28022.992	266773.589	1901638.270	0.87697	0.87688
I(WAR^2)	addition	27715.410	246794.761	1921617.099	0.88619	0.88607
WAR	addition	27553.715	236842.822	1931569.038	0.89078	0.89064
I(OBP^2)	addition	27499.214	233501.814	1934910.046	0.89232	0.89215
OBP	addition	26978.088	204721.791	1963690.069	0.90559	0.90542
RBI	addition	26860.811	198672.985	1969738.875	0.90838	0.90819
PA	addition	26792.269	195180.131	1973231.729	0.90999	0.90979
R	addition	26675.741	189448.919	1978962.941	0.91263	0.91241
HR	addition	26639.402	187631.357	1980780.503	0.91347	0.91323
G	addition	26596.014	185502.153	1982909.706	0.91445	0.91419

Figure 14: ols step both aic using R.

- Variables to add in Model: All.

Stepwise Regression

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	HR	addition	0.895	0.895	904.8620	28205.8850	7.5533
2	R	addition	0.909	0.909	230.6120	27610.4549	7.0230
3	RBI	addition	0.914	0.913	26.3160	27411.3913	6.8536
4	OBP	addition	0.915	0.915	-28.0460	27356.6980	6.8072

Figure 15: ols step both p using R.

- Variables to add in Model: HR, R, RBI, OBP.

Stepwise Regression Results

$$\text{wRC+} = \beta_0 + \beta_1\text{G} + \beta_2\text{PA} + \beta_3\text{HR} + \beta_4\text{R} + \beta_5\text{RBI} + \beta_6\text{BABIP} + \beta_7\text{OBP} + \beta_8\text{wOBA} + \beta_9\text{WAR} + \beta_{10}\text{Off} + \beta_{11}\text{I}(\text{OBP}^2) + \beta_{12}\text{I}(\text{WAR}^2)$$

- The stepwise regression techniques left us with the same model, containing the same variables.

Best Subsets Regression

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.8554	0.8554	0.8553	2771.4793	29503.5606	17874.8406	29522.5147	321391.9290	78.4840	0.0192	0.1447
2	0.8763	0.8762	0.8759	1783.2159	28867.2989	17238.5182	28892.5709	275095.4251	67.1947	0.0164	0.1239
3	0.8859	0.8858	0.8855	1328.9487	28538.0346	16909.1829	28569.6247	253789.9782	62.0058	0.0151	0.1143
4	0.8924	0.8923	0.892	1021.9454	28299.4413	16670.5998	28337.3494	239374.0068	58.4979	0.0143	0.1078
5	0.9005	0.9004	0.9001	638.9046	27980.4114	16352.0013	28024.6375	221387.4585	54.1156	0.0132	0.0998
6	0.9053	0.9052	0.9045	412.3026	27779.1891	16151.1506	27829.7332	210725.4696	51.5220	0.0126	0.0950
7	0.9084	0.9083	0.9076	266.8348	27644.5336	16016.8215	27701.3957	203862.5027	49.8561	0.0122	0.0919
8	0.9105	0.9103	0.9098	173.3684	27555.5725	15928.1298	27618.7526	199435.0312	48.7852	0.0119	0.0899
9	0.9129	0.9127	0.9121	57.5246	27442.4315	15815.4675	27511.9296	193955.6465	47.4564	0.0116	0.0875
10	0.9138	0.9136	0.9129	16.4888	27401.5628	15774.8101	27477.3789	191983.7692	46.9854	0.0115	0.0866
11	0.9148	0.9146	0.9139	-28.0463	27356.6980	15730.2241	27438.8321	189846.6898	46.4737	0.0113	0.0857
12	0.9145	0.9142	0.9135	13.0000	26596.0136	15309.8618	26684.0496	186110.6406	46.9497	0.0118	0.0861

Figure 16: Subsets Regression using R.

- Model 11 and 12 have high R-Squared and low c(p) and AIC, but they have too many variables.
- Model 6 still has high R-Squared and relatively low c(p) and AIC.

Best Subsets Regression

```
                         Best Subsets Regression
----------------------------------------------------------------------
Model Index    Predictors
----------------------------------------------------------------------
     1         wOBA
     2         wOBA Off
     3         wOBA Off I(WAR^2)
     4         OBP wOBA Off I(OBP^2)
     5         OBP wOBA Off I(OBP^2) I(WAR^2)
     6         OBP wOBA WAR Off I(OBP^2) I(WAR^2)
     7         RBI OBP wOBA WAR Off I(OBP^2) I(WAR^2)
     8         PA R RBI OBP wOBA Off I(OBP^2) I(WAR^2)
     9         PA R RBI OBP wOBA WAR Off I(OBP^2) I(WAR^2)
    10         G PA R RBI OBP wOBA WAR Off I(OBP^2) I(WAR^2)
    11         G PA HR R RBI OBP wOBA WAR Off I(OBP^2) I(WAR^2)
    12         G PA HR R RBI BABIP OBP wOBA WAR Off I(OBP^2) I(WAR^2)
----------------------------------------------------------------------
```

Figure 17: Best Subsets Model using R.

- Model 6 is the model we will check further.
- It contains the following variables: OBP, wOBA, WAR, Off, I(OBP^2), $I(WAR^2)$.

Updated Model

$$\text{wRC+} = \beta_0 + \beta_1\text{OBP} + \beta_2\text{wOBA} + \beta_3\text{WAR} + \beta_4\text{Off} + \beta_5\text{I}(\text{OBP}^2) + \beta_6\text{I}(\text{WAR}^2)$$

- Above is our updated model. Now we must check to see if multicollinearity exists in the model by checking the Variance Inflation Factors (VIFs).

Variance Inflation Factors for our Model

	Variables	Tolerance	VIF
1	OBP	0.007845019	127.469417
2	wOBA	0.092240720	10.841199
3	WAR	0.171229677	5.840109
4	Off	0.167569571	5.967671
5	I(OBP^2)	0.008473337	118.017255
6	I(WAR^2)	0.194977525	5.128796

Figure 18: VIFs using R.

- We should not have a variable with a VIF above 10, and we have three.
- We will remove OBP from the model to see how it now affects the VIFs.

Variance Inflation Factors for our New Model

	Variables	Tolerance	VIF
1	wOBA	0.1144321	8.738807
2	WAR	0.1712320	5.840030
3	Off	0.1907936	5.241266
4	I(OBP^2)	0.1345046	7.434692
5	I(WAR^2)	0.1951357	5.124638

Figure 19: VIFs using R.

- We fixed our issue of multicollinearity and created a new model with five variables.
- All we did was remove OBP from the model.

- wRC+ = β_0 + β_1wOBA + β_2WAR + β_3Off + β_4I(OBP^2)$+\beta_5$I(WAR^2)

What our "current" model looks like.

```
Call:
lm(formula = wRCplus ~ wOBA + WAR + Off + I(OBP^2) + I(WAR^2))

Residuals:
    Min      1Q  Median      3Q     Max
-28.898  -4.272   0.396   4.566  35.269

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.549e+01  1.863e+00 -24.424  < 2e-16 ***
wOBA         4.658e+02  9.129e+00  51.030  < 2e-16 ***
WAR          2.173e-01  1.620e-02  13.411  < 2e-16 ***
Off          6.493e-02  2.181e-03  29.778  < 2e-16 ***
I(OBP^2)    -1.071e+02  1.417e+01  -7.564 4.82e-14 ***
I(WAR^2)    -4.079e-03  1.796e-04 -22.706  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.665 on 4091 degrees of freedom
Multiple R-squared:  0.8919,    Adjusted R-squared:  0.8917
F-statistic:  6748 on 5 and 4091 DF,  p-value: < 2.2e-16
```

Figure 20: Summary output for Model using R.

- wRC+ = -45.494 + 465.834$*$wOBA + 0.217318$*$WAR + 0.06493$*$Off - 107.138$*$I(OBP^2)$-0.00408 * I(WAR^2)$

Checking Our Model Assumptions

- 0.) The model is correct.
- 1.) The estimated error is 0 (automatic if least square technique used).
- 2.) The error variance is constant.
- 3.) The errors are normally distributed.
- 4.) The observations are independent.

Checking Our Model Assumptions

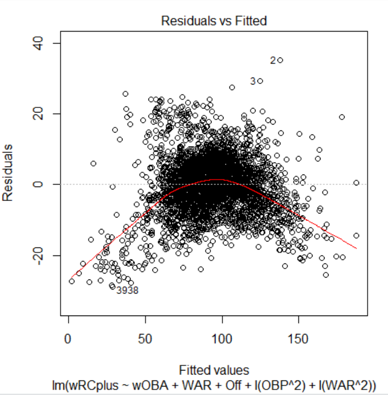

Figure 21: Residuals vs. Fitted plot using R.

- This plot checks assumption 0 and 2.
- There is a clear issue with assumption 0.

Checking Our Model Assumptions

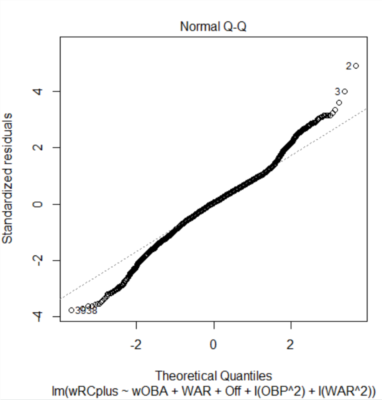

Figure 22: Normal Q-Q plot using R.

- This plot checks assumption 3.
- Not exactly a straight line, but we will accept normality.

Checking Our Model Assumptions

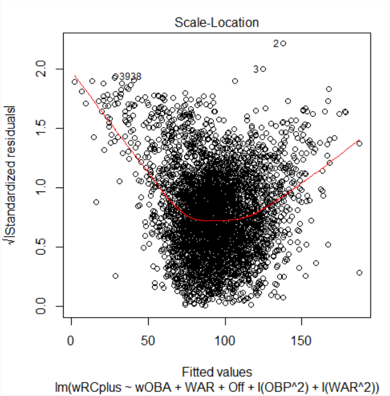

Figure 23: Scale-Location plot using R.

- This plot also checks assumption 2.
- We have an issue with non constant variance.

Checking Our Model Assumptions

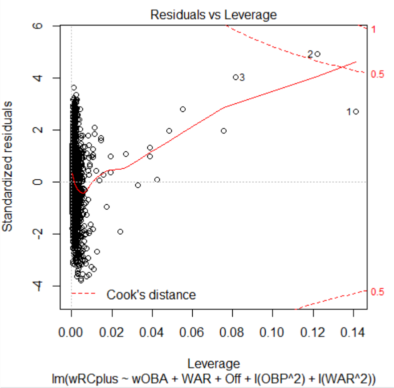

Figure 24: Residuals vs. Leverage plot using R.

- This plot checks for outliers and influential points.
- Some outliers are a few of the best offensive players to play.

Introduction of Interaction Effects

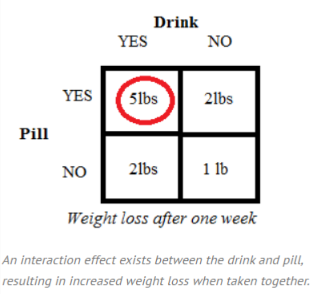

Figure 25: Example of Interaction Effects for weight loss.

- We clearly need to fix our model - so we will look at interaction effects.
- An interaction effect happens when one explanatory variable interacts with another explanatory variable on a response variable.

New Model with Interaction Effects

$$\text{wRC+} = \beta_0 + \beta_1\text{OBP} + \beta_2\text{wOBA} + \beta_3\text{WAR} + \beta_4\text{Off} + \beta_5\text{I}(\text{OBP}^2) + \beta_6\text{I}(\text{WAR}^2) + \beta_7\text{OBP*wOBA} + \beta_8\text{OBP*WAR} + \beta_9\text{OBP*Off} + \beta_{10}\text{wOBA*WAR} + \beta_{11}\text{wOBA*Off} + \beta_{12}\text{WAR*Off}$$

- Above is our new model with interaction effects. Now we will run through best subset regression again to find a smaller model.

Our Final Model

$$\text{wRC+} = \beta_0 + \beta_1\text{wOBA} + \beta_2\text{WAR} + \beta_3\text{OBP*wOBA} + \beta_4\text{wOBA*WAR} + \beta_5\text{wOBA*Off} + \beta_6\text{I}(\text{OBP}^2)$$

What our final model looks like.

```
Call:
lm(formula = wRCplus ~ wOBA + WAR + OBP:wOBA + wOBA:WAR + wOBA:Off +
    I(OBP^2))

Residuals:
    Min      1Q  Median      3Q     Max
-26.789  -3.847   0.031   3.501  54.496

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.292e+02  4.874e+00  -26.51   <2e-16 ***
wOBA         1.008e+03  3.282e+01   30.73   <2e-16 ***
WAR          2.354e+00  7.022e-02   33.52   <2e-16 ***
I(OBP^2)     8.148e+02  5.597e+01   14.56   <2e-16 ***
wOBA:OBP    -1.777e+03  1.083e+02  -16.40   <2e-16 ***
wOBA:WAR    -7.076e+00  2.090e-01  -33.85   <2e-16 ***
wOBA:Off     3.157e-01  6.946e-03   45.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.794 on 4090 degrees of freedom
Multiple R-squared:  0.915,	Adjusted R-squared:  0.9149
F-statistic:  7342 on 6 and 4090 DF,  p-value: < 2.2e-16
```

Figure 26: Summary output for Model using R.

- wRC+ = $-129.2 + 100.8{*}\text{wOBA} + 2.354{*}\text{WAR} + 814.8{*}\text{I}(\text{OBP}^2) - 1777 * wOBA * OBP - 7.076 * wOBA * WAR + 0.3157 * wOBA * Off$

Checking Our Final Model Assumptions

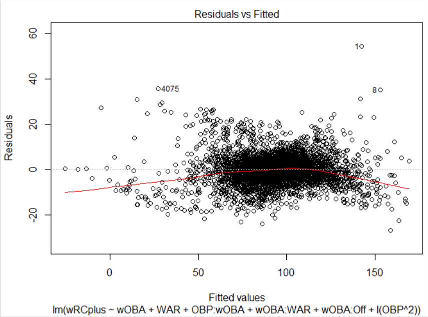

Figure 27: Residuals vs. Fitted plot using R.

- This plot checks assumption that the model is correct and there exists constant variance.
- The model looks clearly better. The red line is much flatter.

Checking Our Final Model Assumptions

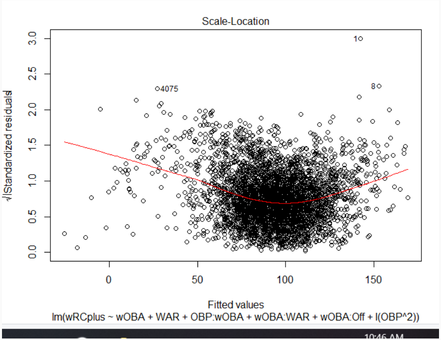

Figure 28: Scale-Location plot using R.

- This plot also checks constant variance.
- The plot shows that we might not have constant variance, but we still accept the model.

Checking Our Final Model Assumptions

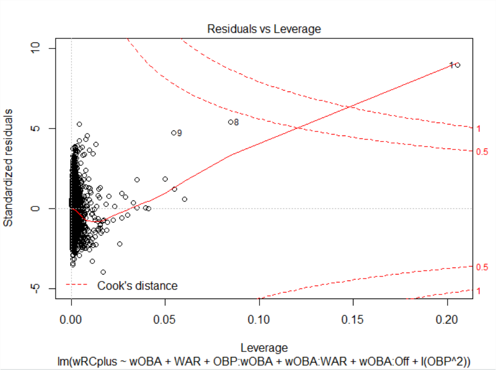

Figure 29: Residuals vs. Leverage plot using R.

- This plot checks for outliers and influential points.
- We have many outliers in our plot. Babe Ruth is an influential point.

Good Examples With Our Model

Names	Actual wRC+	Model wRC+	Difference
Scott Hatteberg	104	104	0
Jay Bruce	106	106	0
Miguel Tejada	106	106	0
Ray Chapman	111	111	0
Joe Mauer	123	123	0

Bad Examples With Our Model

Names	Actual wRC+	Model wRC+	Difference
Ricky Henderson	132	159	27
Gary Sheffield	141	163	22
Ted Williams	188	153	-35
Babe Ruth	197	143	-54
Bobby Mathews	63	27	-36

Relationship of Actual wRC+ vs Model wRC+

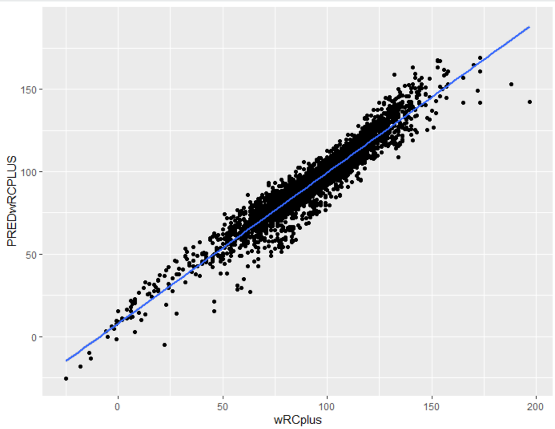

Figure 30: Scatter plot using R.

- R^2 is 0.9150451

Appendix

```r
attach(Summer_2021_Baseball_Research_Data)



mod1.lm <- lm(wRCplus ~ G + PA + HR + R + RBI + BABIP + OBP + wOBA + WAR + Off)
summary(mod1.lm)
library(car)
crPlots(mod1.lm)
#Perhaps second order with OBP and WAR? Log for HR?
mod2.lm <- lm(wRCplus ~ G + PA + HR + R + RBI + BABIP + OBP + wOBA + WAR + Off +
                I(OBP^2) + I(WAR^2))
```

```r
library(olsrr)
ols_step_forward_p(mod2.lm)
#selected variables: HR, R, RBI, OBP
ols_step_forward_aic(mod2.lm)
#selected variables: wOBA, BABIP, Off, I(WAR^2), WAR, I(OBP^2), OBP, RBI, PA, R, HR, G
ols_step_backward_p(mod2.lm)
#selected variables: G + PA + HR + R + RBI + OBP + wOBA + WAR + Off + I(OBP^2) +  I(WAR^2)
ols_step_backward_aic(mod2.lm)
#selected variables: G + PA + HR + R + RBI + BABIP + OBP + wOBA + WAR + Off + I(OBP^2) +  I(WAR^2)
ols_step_both_aic(mod2.lm)
#selected variables: wOBA, BABIP, Off, I(WAR^2), WAR, I(OBP^2), OBP, RBI, PA, R, HR, G
ols_step_both_p(mod2.lm)
#selected variables: HR, R, RBI, OBP
mod2.lm <- lm(wRCplus ~ G + PA + HR + R + RBI + BABIP + OBP + wOBA + WAR + Off + I(OBP^2) + I(WAR^2))
summary(mod2.lm)
k=ols_step_best_subset(mod2.lm)
plot(k)
k
```

```r
mod3.lm <- lm(wRCplus ~ OBP + wOBA + WAR + Off + I(OBP^2) + I(WAR^2))
summary(mod3.lm)
ols_vif_tol(mod3.lm)
plot(mod3.lm)
#Too high of VIF

mod4.lm <- lm(wRCplus ~ wOBA + WAR + Off + I(OBP^2) + I(WAR^2))
summary(mod4.lm)
ols_vif_tol(mod4.lm)
plot(mod4.lm)

mod5.lm <- lm(wRCplus ~ OBP + wOBA + WAR + Off + OBP:wOBA + OBP:WAR + OBP:Off
              + wOBA:WAR + wOBA:Off + WAR:Off+ I(OBP^2) + I(WAR^2))
summary(mod5.lm)
plot(mod5.lm)
k=ols_step_best_subset(mod5.lm)
plot(k)
k

mod6.lm <- lm(wRCplus ~ wOBA + WAR + OBP:wOBA + wOBA:WAR + wOBA:Off + I(OBP^2))
summary(mod6.lm)
plot(mod6.lm)
ols_vif_tol(mod6.lm)
```

```
print(predict(mod6.lm, Summer_2021_Baseball_Research_Data))
PREDWRCPLUS <- predict(mod6.lm, Summer_2021_Baseball_Research_Data)
Summer_2021_Baseball_Research_Data$PredwRCPlus <- PREDWRCPLUS
DIFF <- wRCplus - PREDWRCPLUS
Summer_2021_Baseball_Research_Data$Diff <- DIFF


plot(wRCplus,PREDWRCPLUS)
BaseCor <- Summer_2021_Baseball_Research_Data[,3:23]
cor(BaseCor)

library(ggplot2)
# Basic scatter plot

ggplot(Summer_2021_Baseball_Research_Data, aes(x=wRCplus, y=PREDWRCPLUS)) +
  geom_point() + geom_smooth(method=lm)
(cor(wRCplus,PREDWRCPLUS))^2
```

```r
#calculating leverage of all points

influence(mod6.lm)$hat

Summer_2021_Baseball_Research_Data[which(influence(mod6.lm)$hat> 2*3/25),]

#plotting leverage

plot(influence(mod6.lm)$hat)

#getting all measures of influence together

print(influence.measures(mod6.lm))

#obtaining cooks distance only

cooks.distance(mod6.lm)

#plotting cook's distance

plot(cooks.distance(mod6.lm))
```

```r
#only obtaining observations with high Cook's distance values

Summer_2021_Baseball_Research_Data[which(cooks.distance(mod6.lm) > 1),]

#obtaining DFFITS

dffits(mod6.lm)

#plotting DFFITS

plot(dffits(mod6.lm))


Summer_2021_Baseball_Research_Data[which(abs(dffits(mod6.lm)) > 2*sqrt(3/25)),]

covratio(mod6.lm)
plot(covratio(mod6.lm))
Summer_2021_Baseball_Research_Data[which(abs(covratio(mod6.lm)-1) > 3*3/25),]
```

Sources

Works Cited

- 4, Rob Mains December, et al. "Comparing DRC+, OPS+, and WRC+." Baseball Prospectus, 5 Dec. 2018, www.baseballprospectus.com/news/article/45445/comparing-drc-ops-and-wrc/
- "Baseball Savant: Trending MLB Players, Statcast and Visualizations." Baseballsavant.com, baseballsavant.mlb.com/.
- "FanGraphs Baseball: Baseball Statistics and Analysis." FanGraphs Baseball | Baseball Statistics and Analysis, www.fangraphs.com/.
- Hochman, Benjamin. "Hochman: Esoteric but Useful, Baseball's New Superstat Is WRC+." The Denver Post, The Denver Post, 29 Apr. 2016, www.denverpost.com/2013/07/15/hochman-esoteric-but-useful-baseballs-new-superstat-is-wrc/.

Works Cited

- Kupiec, Ryan. Can Statcast Variables Explain the Variation in Weighted Runs Created plus? scholarship.depauw.edu/cgi/viewcontent.cgi?article=1003context=studentresea
- “The Official Site of Major League Baseball.” MLB.com, www.mlb.com/.
- Sanford, Gavin D. What Raw Statistics Have the Greatest Effect on WRC+ in Major League Baseball in 2017?
- Sharpe, Sam. “An Introduction to Expected Weighted On-Base Average (XwOBA).” Medium, MLB Technology Blog, 24 Sept. 2019
- Stephanie. “Interaction Effect, STATISTICAL Interactions amp; Interacting Variable.” Statistics How To, 12 Oct. 2017, www.statisticshowto.com/interaction-effect-interacting-variable/.

Questions?