# Predicting Career Wins Above Replacement from Rookie Stats in Baseball


By: Jack Blumenstein, Josh Braverman, Lekh Murthy and Wilson Wendt


Wharton Moneyball Academy 2022, Sports Analytics Student Research Journal

# Background:
# The Esteban German Dilemma

Esteban German had a good rookie season, which would seem to lead to later success.

| Season | Team | Level | Age | G | PA | HR | R | RBI | SB | BB% | K% | ISO | BABIP | AVG | OBP | SLG | wOBA | xwOBA | wRC+ | BsR | Off | Def | WAR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2006 | KCR | MLB | 28 | 106 | 331 | 3 | 44 | 34 | 7 | 12.1% | 14.8% | .133 | .388 | .326 | .422 | .459 | .390 | | 136 | -1.3 | 14.0 | -9.3 | 1.6 |

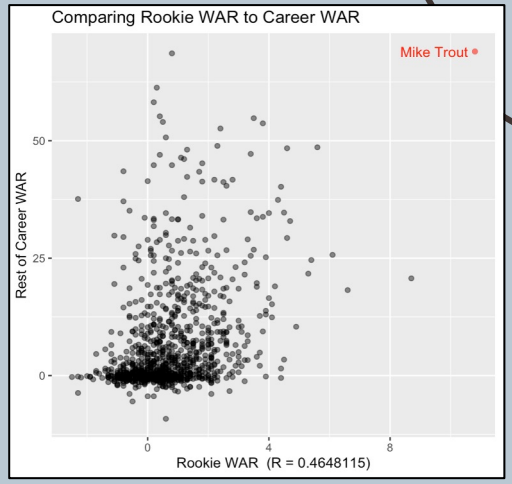
However, over the rest of his career he compiled a lower WAR (1.3) in 5 seasons then he did in his rookie year (1.6). Why??

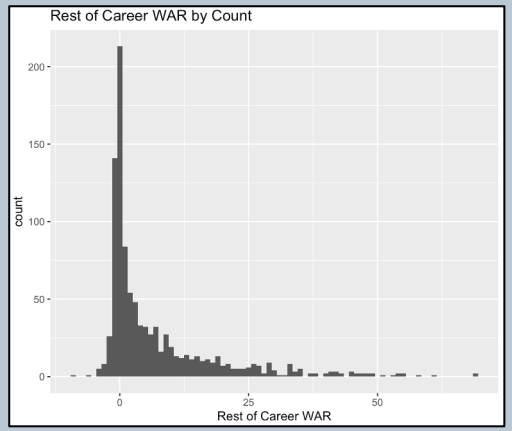
| | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2007 | KCR | MLB | 29 | 121 | 405 | 4 | 49 | 37 | 11 | 10.6% | 14.8% | .112 | .307 | .264 | .351 | .376 | .327 | | 94 | -0.7 | -3.5 | 1.1 | 1.1 |
| 2008 | KCR | MLB | 30 | 89 | 242 | 0 | 30 | 22 | 7 | 7.4% | 17.4% | .093 | .299 | .245 | .303 | .338 | .285 | | 66 | 0.2 | -9.8 | -1.6 | -0.3 |
| 2009 | TEX | MLB | 31 | 19 | 50 | 0 | 9 | 4 | 1 | 8.0% | 14.0% | .087 | .359 | .304 | .360 | .391 | .336 | | 96 | 1.0 | 0.7 | -1.4 | 0.1 |
| 2010 | TEX | MLB | 32 | 13 | 16 | 0 | 5 | 1 | 4 | 18.8% | 12.5% | .000 | .273 | .231 | .375 | .231 | .299 | | 76 | 1.1 | 0.6 | -0.5 | 0.1 |
| 2011 | TEX | MLB | 33 | 11 | 13 | 1 | 6 | 4 | 1 | 7.7% | 7.7% | .364 | .400 | .455 | .462 | .818 | .517 | | 227 | 0.2 | 2.1 | 0.4 | 0.3 |

# Rookie WAR vs. Career WAR

- There is no clear trend between Rookie and Career WAR
- Many points are clustered near the origin with random outliers.
- Only 25% of players in the dataset get more than 10 WAR after their rookie season

Rookie WAR is not a good predictor for Rest of Career WAR, so...



# Multivariate Linear Regression

We found specific stats that helped us understand what commonalities rookies with future success shared:

We found:

- Power (ISO)
- Strikeouts (K%)
- Stolen Bases (SB)
- Quality of Contact (HH%/SH%)
- Defensive Ability (Def/PA)
- Age (Years)
- Team Value (PA/G)

Notable Exclusions:

- Walks
- Position

Logarithmic variables:

- Log(Quality of Contact)
- Log(Age)

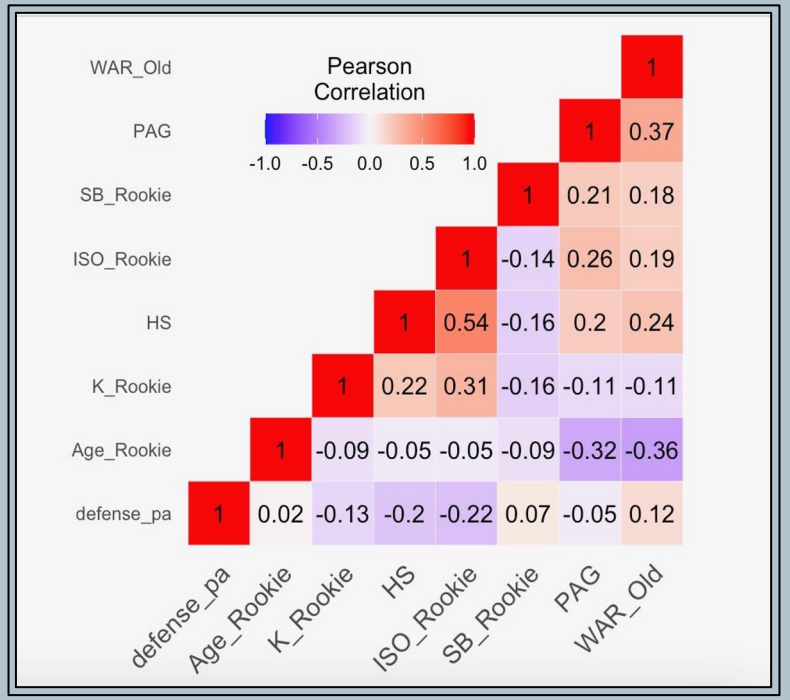
# Data Scraping/Cleaning

- We obtained our data from FanGraphs

Filters

- Rookies vs. Rest of Career
- 2003-2015 vs. 2016-2022
- Rookie Status/Rookie of the Year Voting
- Added the variables that we used in the regression
- Logarithmic Relationships
- Changed variables into rates

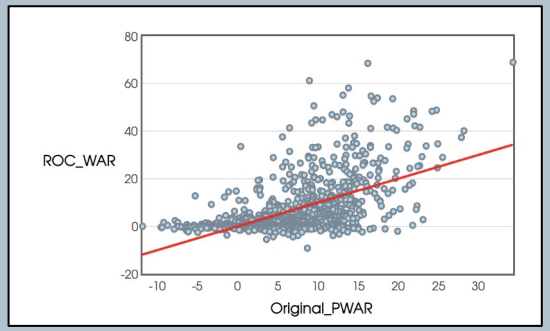
# Correlation Table

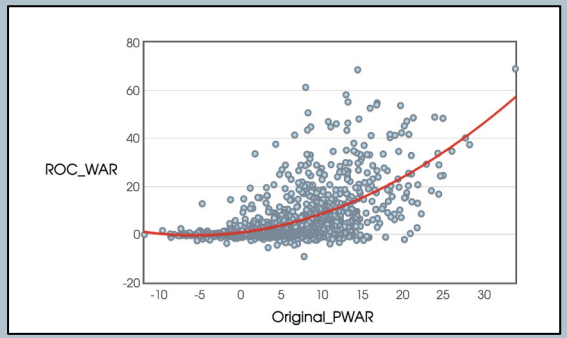


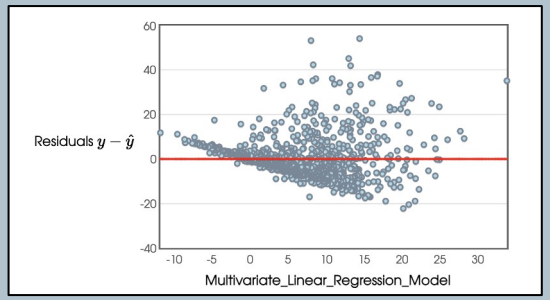

$$0.72111+0.34493*(125.5398-(40.4491*log(AgeRookie))+(0.2039*SBRookie)+(86.1499*defensepa)+(26.3500*ISORookie)-(31.2288*KRookie)+(5.6711*log(HS))+(2.9340*PAG))+0.04214*(125.5398-(40.4491*log(AgeRookie))+(0.2039*SBRookie)+(86.1499*defensepa)+(26.3500*ISORookie)-(31.2288*KRookie)+(5.6711*log(HS))+(2.9340*PAG))^2$$

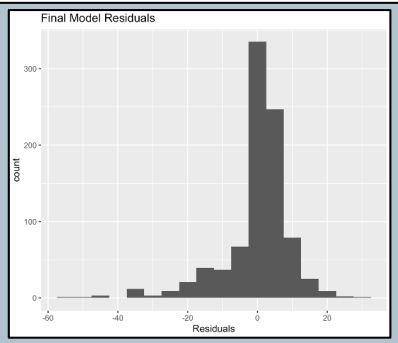
# Results

Final Model Correlation: r=0.596, RMSE=9.55 WAR, R.E.=0.197



# Predictions

| Name | PWAR |
|---|---|
| Ronald Acuna Jr. | 36.76367442 |
| Trea Turner | 32.97564397 |
| Ozzie Albies | 31.73422701 |
| Julio Rodriguez | 30.85357606 |
| Wander Franco | 28.42303068 |
| Cody Bellinger | 27.85403839 |
| Bobby Witt Jr. | 26.67330673 |
| Juan Soto | 26.59037346 |
| Yordan Alvarez | 26.17991961 |
| Fernando Tatis Jr. | 25.72066506 |
| Michael Harris II | 25.3428497 |
| Gary Sanchez | 23.52713142 |
| Tommy Edman | 23.21538293 |
| Trevor Story | 22.868162 |
| Corey Seager | 22.61155529 |
| Bo Bichette | 22.53746863 |

# Validation

- Dataset is trained on 2003-2015 but still works great on 2016-2018 data even with their careers not being over yet

- 2016
  - 0.621 r
  - 0.216 R.E.
- 2017
  - 0.696 r
  - 0.282 R.E.
- 2018
  - 0.581 r
  - 0.185 R.E.


# Randy Arozarena vs. Alejandro Kirk

- In the 2021 ROTY voting, Arozarena won while Kirk didn't receive a vote.
- This year, Kirk has had a breakout season, while Arozarena has regressed.
- Awards based off of one season can be misleading.

Randy Arozarena, TB: 22 (first-place votes), 4 (2nd), 2 (3rd) -- 124 points
Luis Garcia, HOU: 2 (1st), 15 (2nd), 8 (3rd) -- 63 points
Wander Franco, TB: 2 (1st), 5 (2nd), 5 (3rd) -- 30 points
Adolis García, TEX: 3 (1st), 1 (2nd), 9 (3rd) -- 27 points
Emmanuel Clase, CLE: 1 (1st), 2 (2nd) -- 11 points
Ryan Mountcastle, BAL: 2 (2nd), 4 (3rd) --10 points
Shane McClanahan, TB: 1 (2nd) -- 3 points
Alek Manoah, TOR: 2 (3rd) -- 2 points

| Name | PWAR |
|---|---|
| Alejandro Kirk | 19.12092533 |
| Randy Arozarena | 7.419731217 |

# Limitations

- A few of the players in the training dataset have not yet finished their careers
- Minor league success isn't taken into account
- Steroid era limits data
- No Statcast data
- Injuries


# Conclusion

- Career WAR can be projected well from rookie metrics using our model
- Some players that have a strong start to their careers usually regress because of unsustainable measures in their underlying statistics.
- **Our Model actually predicted Esteban German well because of poor underlying statistics in his rookie season**

# References

FanGraphs.com

https://www.remove.bg/upload

https://slidesgo.com

https://www.mlb.com/news/2021-mlb-rookie-of-the-year-voting-results

https://stats.blue/

# Questions?