

Topic: Detection of Anomalous Baseball Performance For Detection of Possible Steroid Use

1. Description of the particular problem within the selected data mining topic to be addressed in this project: Performance-enhancing drugs have come to the public eye in recent decades in several professional sports league. Major League Baseball in particular has had steroid abuse scandals since the turn of the century that were sufficiently extensive as to merit attention by the United States Congress.

By identifying players who are having anomalously successful seasons, we can refine the process of steroid detection. Over a thousand players participate in Major League Baseball every season, making data mining a useful tool in detection. This will include players who are having anomalously good seasons relative to their own historic performance, players whose career is defined by anomalous success, and players whose age has less impact than expected. Further steps in this process are likely best left to investigators, as it can be difficult to differentiate between a player who has boosted their performance using illegal substances and one who is enjoying an extended run of good fortune.

2. Description of the approach used in this project to tackle the above problem: Many important baseball statistics can be treated as normally distributed with appropriate data pre-processing. Therefore, by using z-score as an anomaly score, we can assess the likelihood of seeing particular results based on larger data sets. Specifically, we will use the robust z-score that incorporates median rather than mean, as we anticipate outliers. These sets will either be the entirety of a player's career, if focused on fluctuations in an individual's performance, and aggregate performance across the league if focused on overall steroid use.

3. Dataset Name: Baseball-Reference.com's WAR data

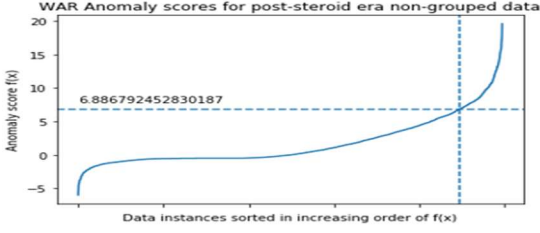
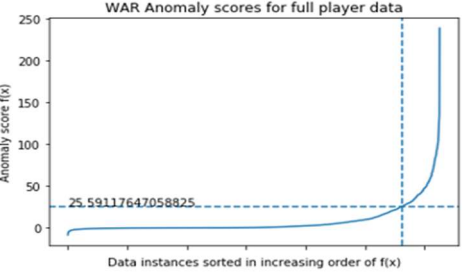
4. Where found: <https://www.baseball-reference.com/about/>

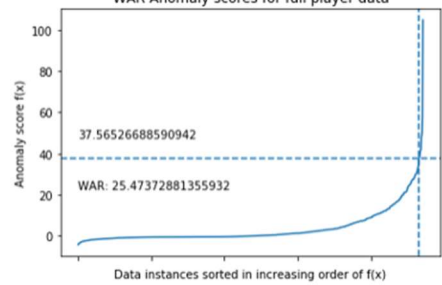
5. Dataset Description: A vast amount of data is tracked on the baseball box score and in other capacities to evaluate player performance, much of which is publicly available. Both cited datasets have one data entry per player per year since 1871, totaling about 105,000 entries, each containing a total of 71 attributes of player and team identification (all strings), and performance attributes (all either integers or floating-point numbers). Most attributes were deleted, keeping those needed for identification (player names), relevant contextual statistics (age, games played) and the key relevant statistic, Wins Above Replacements (WAR) as well as WAR solely from offense, excluding defense.

6. Initial data preprocessing, if any: The steroid era is taken to be 1993-2002. To provide a balanced context, analysis was be done on the years 1983-2012. The attributes not enumerated above were deleted. Any season with less than 81 games played was excluded to avoid skew. Baseball statistics are rigorously kept and validated, meaning that no missing or erroneous values occurred.

7, Three Guiding Questions about the dataset domain:

1. Does overall performance in years of the steroid era show anomalously high levels relative to surrounding years?
2. Do individual careers in the years of the steroid era show anomalously high performance relative to surrounding years?
3. Is performance in the steroid era anomalously high for older players?

Summary of Experiments. All performed with Python. Mining technique: anomaly detection				
Guid. Question	Pre-process	Parameter	Eval.: F(x) and WAR at cut-off point	Observations about experiment Observations about visualization Interpretation results
1	Only as above	Top 10% are anomalies	Full: 5.65, 3.94 Pre-steroid: 4.40, 4.04 Steroid: 6.12, 3.92 Post-steroid: 6.89, 3.9	<p>The below visualization of instances against anomaly score is typical for this guiding question.</p>  <p>For the pre-steroid data set, it is less anomalous to have a higher score. The other three categories are broadly comparable.</p>
1	Only as above	Top 10	Full: 15.01, 9.84 Pre-steroid: 10.19, 8.55 Steroid: 14.91, 9.11 Post-steroid: 16.02, 8.74	The pre-steroid data set has a lower anomaly score for a broadly comparable WAR.
1	As above, offense only	Top 10% are anomalies	Full: 5.11, 3.69 Pre-steroid: 3.98, 3.72 Steroid: 5.94, 3.73 Post-steroid: 6.09, 3.63	The pre-steroid data set has a lower anomaly score than the other categories, and all have a similar WAR.
1	As above, offense only	Top 10 are anomalies	Full: 13.44, 8.94 Pre-steroid: 9.23, 7.77 Steroid: 14.87, 8.73 Post-steroid: 13.67, 7.72	The full dataset is very similar to the Steroid data set.
2	Grouped by individual	Top 10% are anomalies	Full: 25.59, 17.65 Pre-steroid: 12.55, 15.08 Steroid: 21.97, 13.6 Post-steroid: 24.66, 13.82	<p>The below visualization of instances against anomaly score is typical for this guiding question.</p>  <p>The full dataset has the highest anomaly score for the highest WAR. The visualization shows a massive outlier with an anomaly score of around 250. This is Barry Bonds, a known steroid user who set numerous records</p>
2	Grouped by	Top 10 are anomalies	Full: 105.51, 71.99 Pre-steroid: 37.09, 55.07 Steroid: 76.59, 46.91	The anomaly score for the full dataset is very large, and all anomaly scores are higher than seen previously.

	individual		Post-steroid: 69.20, 38.54	
2	Grouped by individual, offense only	Top 10% are anomalies	Full: 23.71, 16.65 Pre-steroid: 11.56, 13.78 Steroid: 21.88, 13.16 Post-steroid: 22.92, 12.57	The anomaly score for the pre-steroid data set is still smaller than the others for comparable WAR.
2	Grouped by individual, offense only	Top 10 are anomalies	Full: 107.55, 74.5 Pre-steroid: 33.86, 48.75 Steroid: 77.33, 45.88 Post-steroid: 67.98, 36.90	The anomaly score for the full dataset is very large, and all anomaly scores are higher than seen previously.
3	Grouped by individual, Age > 30	Top 10% are anomalies	Full: 13.17, 9.12 Pre-steroid: 8.01, 7.44 Steroid: 13.02, 8.83 Post-steroid: 14.94, 7.43	<p>The below visualization of instances against anomaly score is typical for this guiding question.</p>  <p>Similar trends as in the rest of the guiding questions. There is a massive anomaly – this is Barry Bonds again.</p>
3	Grouped by individual, Age > 30	Top 10 are anomalies	Full: 37.56, 25.47 Pre-steroid: 17.23, 20.83 Steroid: 24.50, 16.41 Post-steroid: 30.13, 14.95	Post-steroid era has notable smaller WAR than other time periods.
3	Grouped by individual, Age > 30, offense only	Top 10% are anomalies	Full: 12.96, 9.58 Pre-steroid: 8.65, 8.42 Steroid: 12.22, 9.07 Post-steroid: 16.63, 7.74	Post-steroid era has significantly higher anomaly score and lower WAR.
3	Grouped by individual, Age > 30, offense only	Top 10 are anomalies	Full: 36.57, 26.34 Pre-steroid: 15.40, 22.54 Steroid: 23.11, 16.74 Post-steroid: 41.07, 18.98	Post-steroid era has significantly higher anomaly score and higher WAR.

Analysis of Results: The results are reproduced below with the same format of <Anomaly score, WAR>

No grouping				
Data set	Full	Pre-steroid	Steroid	Post-steroid
Top 10% are anomalies	5.65, 3.94	4.40, 4.04	6.12, 3.92	6.89, 3.9
Top 10% are anomalies, offense-only	5.11, 3.69	3.98, 3.72	5.94, 3.73	6.09, 3.63
Top 10 are anomalies	15.01, 9.84	10.19, 8.55	14.91, 9.11	16.02, 8.74
Top 10 are anomalies, offense only	13.44, 8.94	9.23, 7.77	14.87, 8.73	13.67, 7.72
Grouped by Individual, Top 10% are anomalies				
Data set	Full	Pre-steroid	Steroid	Post-steroid
All	25.59, 17.65	12.55, 15.08	21.97, 13.6	14.66, 13.82
All, offense-only	23.71, 16.65	11.56, 13.78	21.88, 13.16	22.92, 12.57
Age > 30	13.17, 9.12	8.01, 7.44	13.02, 8.83	14.94, 7.43
Age > 30, offense only	12.96, 9.58	8.65, 8.42	12.22, 9.07	16.63, 7.74
Grouped by Individual, top 10 are anomalies				
Data set	Full	Pre-steroid	Steroid	Post-steroid
All	105.51, 71.99	37.09, 55.07	76.59, 46.9	69.20, 38.54
All, offense-only	107.55, 74.5	33.86, 48.75	77.33, 45.88	67.98, 36.90
Age > 30	37.56, 25.47	17.23, 20.83	24.50, 16.41	30.13, 14.95
Age > 30, offense only	36.57, 26.34	15.40, 22.54	23.11, 16.74	41.07, 18.98

We can see that when the data is not grouped by individual players, the steroid and post-steroid eras have similar results, and the anomaly score shows the pre-steroid era to have similar high-end performance that is less anomalous compared to the rest of the dataset, suggesting a higher overall level of play. This trend generally holds true across all guiding questions, and suggests that, surprisingly, the steroid era led to an overall reduction in WAR in terms of overall performance, individual careers, and players over the age of 32.

The z-scores for the top 10 performers when grouped by individual are the most interesting results, as aside from the pre-steroid era they are massive compared to the other anomaly scores shown. This suggests that the top 10 performers are huge outliers, and does support the notion that steroids had an effect and continued to do so even after the steroid era ended.

The visualizations, shown again below, show the same trends of some noticeably negative performances, many mediocre performances, and a few large outliers, particularly Mr. Bonds.



Summary of what you learned in this project:

For a simple $f(x)$, anomaly scores are relatively easy to calculate in Python and can reveal a wealth of information about a desired topic, if the right questions are asked. In the domain, I was surprised by the results, as they generally contravene common wisdom that steroids should boost performance. In general, though, the results suggest that steroid usage hasn't meaningfully changed since the end of the "steroid era" as baseball commentators like to pretend.