Outlier Identification in Count Data Using Variance Difference

Alexander Pelaez, Hofstra University, Hempstead, NY, 11549, Alexander.Pelaez@hofstra.edu
Elaine Winston, Hofstra University, Hempstead, NY, 11549, Elaine.R.Winston@hofstra.edu
Nooshin Nejati, 5 Element Analytics, LLC, Merrick, NY 11566, nnejati@5eanalytics.com
Jiangbing Zhu, 5 Element Analytics, LLC, Merrick, NY 11566, jzhu@5eanalytics.com

Introduction

This article discusses the use of statistical methods to improve competitive advantage in team sports. The focus for this paper is on baseball and more particularly standard and popular metrics used in determining player performance. Baseball's popularity is steadily increasing around the world as more countries play the game. The United States, Japan, South Korea, and many Latin American countries rank baseball as the number one sport. Increasing team revenue has led the assignment of large contracts to individual team players; there is continuous pressure on the team management to find and maintain talent. Many teams have turned to Sabermetrics, a term used for a series of special statistics used in baseball. Traditionalists dislike the use of analytics for many of the decisions in the game, however, the professional teams apply analytics to gain better team positioning. An increased demand for more relevant and accurate analytics spurs research in the area among teams, however, there hasn't been a corresponding increase of academic research on the subject. Baseball statistics offer a wealth of opportunity to researchers.

In this article, we propose an examination of a single statistic, the Earned Run Average, ERA, a comparative measure for pitching success. Identifying a single bad performance, e.g. outlier, can provide a more representative measure of actual performance. Therefore, methods for defining outliers are imperative. Specifically, we provide a few alternatives for the removal of outliers, and propose a method by which to test its effectiveness.

Literature Review

Earned Run Average

Statistics, plays an integral role in the game of baseball. Through the years, measures for hitters such as batting average (AVG), On Base Percentage (OBP), and Slugging Percentage (SLG) have been used to rank hitters. Pitchers' ranking relies on total wins (W), strikeout to walk ratio (SO/BB), and Earned Run Average (ERA) rank hitters. In the 1970's, Bill James, argued that the statistics employed do not accurately reflect the nature of the game (Puerzer, 2002). The work by Bill James and others in the field led to the development of metrics such as OBP, "On Base Percentage", to include ways a hitter can get on base without a hit (Albert, 2010), and SLG, "Slugging Percentage", to adequately weight hits that have greater value such as Home Runs (Gould, 2002).

For a team to be successful, it must win, and pitchers are a critical component for a team's victory. Assessing a measure of success for pitchers should be related to wins. However, a game may have multiple pitchers, but only one pitcher can get a "win". Therefore, the Earned Run Average has traditionally been used as the measure of successful performance. Earned Run Average is calculated as the number of Earned Runs divided by the number of innings pitched, times 9. It can be interpreted as the estimated number of earned runs a pitcher would yield in a nine inning, or complete, game.

*ERA = 9 * Number of earned runs / Number of innings pitched*

The ERA measure has been examined by researchers as a useful analytic for examining the effect of pay for performance (Sommers and Quinten, 1982), Discrimination (Andersen and LaCroix, 1991), and a comparative measure to psychological performance (Smith and Christensen, 1995). The flaw in the Earned Run Average however, is that it is not independent of team performance, nor is it a measure of the contribution made by the pitcher (Scully, 1972).

Sabermetrics attempts to alleviate the problems with the use of ERA by introducing a number of alternative statistics such as Walks and Hits per Innings Pitched (WHIP), Defense-Independent ERA (dERA), Adjusted Earned Run Average (ERA+) and Fielding Independent Pitching (FIP). The numerous available statistics suggests the difficulty in providing any single measure of performance for pitchers, however, ERA is still a widely used and quoted measure of performance.

The flaws of the ERA are fairly well documented. The arguments against ERA are rooted in the ERA's sensitivity to the negative performance of the defense, or even by how scorers attribute an earned run. Further, managers may leave a pitcher in longer to provide more experience for a young pitcher, thereby increasing the probability of more runs, earned or otherwise, which in turn increases the ERA. In addition, a single bad performance by either the pitcher or their defense could inflate the pitcher's overall ERA, and therefore, not be a respective measure of actual performance. Therefore, identifying these single performances that could possibly skew a pitcher's true overall performance could be useful, and thus we note these performances as outliers.

A number of methods are prescribed to identify outliers in a data set. One method would be to identify outliers as lying outside 3 standard deviations from the mean; however, since the distribution of earned runs is not normal, this may not be the most effective method. A common method of outlier detection is the method mentioned by Tukey (1977) whereby one uses the IQR as the principle point of reference. According to this method, observations that are 1.5 times beyond the IQR are considered mild outliers, "inside the fence" and those that are 3 times the IQR are considered extreme outliers, "outside the fence". While Tukey makes no

assumption on the distribution, it does assume the data is continuous (Tukey, 1977; Hubert and Vandervieren, 2008). Alternative, more robust measures have been made for normal distributions such as Grubbs Test (Grubbs, 1969), and an alternative for asymmetric distributions (Carling, 2000).

Grubbs (1969) provides the rationale that an outlier naturally will be a measure of the distance to the mean, and thus extreme observations are those with larger distances. An alternative method, known as the alternative box plot method (Iglewicz and Banerjee, 2007) suggest that the sample and distribution will affect the multiplier proposed by Tukey (1977). These authors noted that their method works when sample sizes are large enough. Yet, they proposed alternative multipliers for smaller data sets. Other methods such as the Variance Shift Outlier Model (VSOM) have been proposed to identify outliers in linear models (Gumedze et al., 2010, Cook and Weisberg, 1982). Research suggests that using the variance change for identification of outliers in a univariate dataset is useful and promising (Tsay, 1988).

In this article, we attempt to identify the proper distribution of the earned run, which is count based data. Furthermore, in order to provide a more representative measure of actual performance, we attempt using a popular method for outlier detection. The purpose of this paper is only to identify right tailed outliers, since it is more likely for a professional pitcher to earn zero runs in a game than giving up 10 runs in a game. Finally, we propose an alternatives method for the removal of outliers against the popular method as a comparison.

Methodology

Distribution of Earned Runs

It is important to understand the distribution of Earned Runs at the game level. It is possible that Earned Runs may follow a Poisson Distribution, Zero-Inflated Poisson Distribution, or a Negative Binomial Distribution, since the data is count based. Dolinar (2014) provided evidence that the actual distribution of runs per game might be a negative binomial distribution. Assuming that the number of Earned Runs follows any of the above distributions, then the ERA statistic used in baseball is clearly flawed as a measure of performance, since it doesn't take into account the distribution.

Using sample data obtained for the top ten pitchers in the National League, we had 301 observations, i.e. games pitched, with each pitcher pitching at least 28 games. Therefore, we had a reasonable sample of pitchers for this pilot. The ERA for these pitchers ranged from 2.33 to 3.51. A closer examination of the total Earned Runs for each pitcher showed a variance higher than the mean overall ($\mu = 62.4$, $\sigma^2 = 125.6$). Similarly, the variance exceeded the mean for earned runs per game ($\mu = 2.07$, $\sigma^2 = 3.41$).

Since the measure of Earned Run Average is used to assess a pitcher's performance, we decided to focus on the earned runs per game. The goal is to determine if the performance of a

single game should be measured, and possibly removed, as an outlier. Using the data available, we analyzed the earned runs and compared them to the described theoretical distributions (Table 1).

ER	Actual	E(X) - Poisson	E(X) - zip	E(X) - NBin
0	0.213	0.124	0.269	0.211
1	0.249	0.259	0.149	0.260
2	0.203	0.270	0.196	0.210
3	0.130	0.188	0.172	0.140
4	0.096	0.098	0.113	0.084
5	0.060	0.041	0.060	0.046
6	0.020	0.014	0.026	0.024
7	0.017	0.004	0.010	0.012
8	0.010	0.001	0.003	0.006
9	0.003	0.000	0.001	0.003
X^2		-	16.57	43.31
LogLik		-586.29	-572.92 (p<.01)	-564.64 (p<.01)

Table 1: Distribution of Earned Runs Compared to Theoretical Distributions

Examining the χ^2 against the actuals, the negative binomial distribution appears to the best fit. While there might have been a possible inflated number of zeros, the use of the zero inflated Poisson doesn't necessarily apply. The application of zero-inflated Poisson is most notable when observations or respondents in the dataset didn't have any opportunity for a treatment, such as functional decline in aging (Byers et al., 2003), insurance claims (Bouchere et al., 2008), length of hospital stay with sepsis (Yang et al., 2010).

Having determined the most appropriate distribution, we next approach the issue of outliers. Identifying outliers is critical to ensuring proper predictions or estimates. Outliers are considered observations that deviate significantly from other observations, or raise concerns that the observations were a result of a different mechanism (Grubbs, 1950; Hawkins, 1980). Removal of proper outliers will provide different measures of performance in many cases (Grubbs, 1950). The impact of these removed observations should therefore yield results that are more indicative of true performance.

Specifically, our aim is to identify an approach for the removal of outliers within a univariate data set for the purposes of using the standard metrics within baseball data. Combining all of the data of earned runs we examine what data points would be potential outliers. First, we apply the Tukey method for outlier identification for the right tail. Based on our complete dataset, the inside fence, i.e. Q3+1.5(Q3-Q1) is 6, and the outside fence is Q3 + 3(Q3-Q1), is 9. The method identified 9 points out of the 301 observations for the ten pitchers as being possible outliers.

Further, we propose another method which examines the variance difference. The VSOM model described above for linear models removes observations with inflated variance (Gumedze et al., 2010, Cook and Weisberg, 1982). Our objective is to narrow the range of outliers between the inner fence, which maybe too conservative, and the outer fence, which may not pick up outliers effectively. Since negative binomial distribution is count data, the raw difference could be as little as 1 or 2, but the impact of more accurately identifying the outliers are critical to the Earned Run Average. Specifically, we propose measuring the difference in variance attributed to the removal of high right side outliers. While, it may be possible to remove left side outliers, we do not examine this particular point.

According to Cochran's theorem, under a normal distribution, the sample variance follows a chi-squared distribution. In addition, it has been shown that the chi-squared distribution is a reasonable approximation for the index of dispersion (Loukas and Kemp, 1986). Thus, we combine this notion of comparing the updated variances to a Chi-squared distribution. Using this information, we calculate the right tail of the chi-squared distribution to identify which variances are beyond a certain point of the Chi-Square.

In our case, 301 observations, the variance for the data set was 3.41 (s^2 = 3.41). We then calculated the variance of the data when each observation is removed and calculate the difference. We compare this to a Chi-Square distribution with degrees of freedom equal to the mean of the differences. Using this method, we calculate the possible outliers and the number of outliers removed whose differences are greater than the calculated Chi-square statistic.

	97.5%	98%	98.5%	99.0%

χ^2	.0095	.0252	.0673	.1852
Outlier Range	2-9	6-9	7-9	Inf / Inf
# of outliers	94	15	9	0

Table 2: Prospective outlier values

Table 2 contains the results of prospective outlier values for the entire dataset. At 98.5% level, we remove the same number of outliers as the boxplot method, i.e. Tukey inner fence method. Since our aim is to identify possible outliers for each pitcher, we turn our attention to the removal of outliers for a given pitcher.

By using the same method, we attempt to identify outliers for possible removal for each individual pitcher. However, due to the smaller sample size, i.e. the number of games per pitcher is approximately 30. After a few examinations, it became clear that we needed to extend the significance level down to 95%. When the confidence level is reduced we end up with the following table.

Pitcher ID	95%	96%	97%	98%	99%	Tukey Outliers
1	1	0	0	0	0	1
2	1	1	0	0	0	3
3	1	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	1	1	0	0	0	1
7	1	0	0	0	0	3
8	1	1	0	0	0	2

9	0	0	0	0	0	0
10	1	0	0	0	0	1

Table 3: Number of Possible Outliers at each level

The numbers in the columns of Table 3 indicate the quantity of possible outliers at the given level. The last column contains the number of outliers identified using the Tukey, inner fence, method. It can be shown from the table above that the number of outliers identified by the proposed method matches in 6 cases. In only one case did the method, propose an outlier, when the Tukey method did not, otherwise the method was more conservative in the outlier estimation. Overall, for each individual pitcher the Tukey method identified 11 outliers, while using our proposed method, only 7 outliers were identified.

Conclusion

Our preliminary analysis shows that it is possible to identify outliers using the above proposed method. Since the Tukey multiplier of 1.5 and 3.0 are given as reasonable measures, our aim was to provide a more robust identification, that could be used for count data, which is a goal of this method. We believe this method shows promise for use in count data, specifically for baseball data, but clearly with applications in other areas such as healthcare hospital admissions or insurance claim analysis.

It is believed that the confidence level for the Chi-square statistic is related to the size of the sample. When all the data is combined, it was shown that the confidence level could be more conservative, but when the sample drops to around 30, we find that the significance level drops to around 95%. We believe this will probably be a reasonable estimate for small samples. Further, the confidence level could be proportional to the sample size.

Further research with larger sets and subsets are proposed in order to more fully examine the effect of this. While other research has been done in the identification of outliers, our purpose is to provide a simple approach, similar to Tukey, using a method based on a statistical distribution such as the Chi-Squared. In addition, this research should be extended to other count data sets to examine the differences between data from different distributions. Although the main focus of this paper was to detect right tailed outliers, further analysis should be conducted to identify left tailed outliers in the future. This analysis can be extended to the identification of false negatives and false positives on both left tail and right tail for more accuracy.

Acknowledgement
The authors would like to acknowledge the two blind reviewers for their helpful comments in the final version of the paper. We would also like to thank Tyler Levine of the Long Island Ducks Professional Baseball Team and the people at Long Island Baseball in Bellmore, NY for their assistance in the development of this paper.

References

Albert, J. (2010). Sabermetrics: The past, the present, and the future. *Mathematics and sports*, (43), 15.

Andersen, T., & Croix, S. J. (1991). Customer racial discrimination in major league baseball. *Economic Inquiry*, *29*(4), 665-677

Banerjee, S., & Iglewicz, B. (2007). A simple univariate outlier identification procedure designed for large samples. *Communications in Statistics—Simulation and Computation®*, *36*(2), 249-263.

Boucher, J. P., Denuit, M., & Guillén, M. (2008). Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions. *Variance*, *2*(1), 135-162.

Byers, A. L., Allore, H., Gill, T. M., & Peduzzi, P. N. (2003). Application of negative binomial modeling for discrete outcomes: a case study in aging research. *Journal of clinical epidemiology*, *56*(6), 559-564.

Carling, K., 2000. Resistant outlier rules and the non-Gaussian case. Comput. Statist. Data Anal. 33 (3), 249–258. Cleveland, W.S., 1985.

Cook, R.D., Weisberg, S., 1982. Residuals and Influence in Regression. Chapman and Hall, New York.

Dolinar, S. (2014, Sept). Run Distribution Using the Negative Binomial Distribution. Retrieved From https://www.fangraphs.com/community/run-distribution-using-the-negative-binomial-distribution/ , Nov 10, 2017.

Last, F. M. (Year, Month Date Published). Article title. Retrieved from URL.

Gould, S. J. (2010). *Triumph and tragedy in Mudville: A lifelong passion for baseball*. WW Norton & Company.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 27-58.

Grubbs, F. E. (1969). Procedures for detecting outlying

Gumedze, F. N., Welham, S. J., Gogel, B. J., & Thompson, R. (2010). A variance shift model for detection of outliers in the linear mixed model. *Computational Statistics & Data Analysis*, *54*(9), 2128-2144.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, *52*(12), 5186-5201.

Loukas, S., & Kemp, C. D. (1986). The index of dispersion test for the bivariate Poisson distribution. *Biometrics*, 941-948.

Puerzer, R. J. (2002). From scientific baseball to sabermetrics: Professional baseball as a reflection of engineering and management in society. *NINE: A Journal of Baseball History and Culture*, *11*(1), 34-48.

Scully, G. W. (1974). Pay and performance in major league baseball. *The American Economic Review*, *64*(6), 915-930.

Smith, R. E., & Christensen, D. S. (1995). Psychological skills as predictors of performance and survival in professional baseball. *Journal of Sport and Exercise Psychology*, *17*(4), 399-415.

Sommers, P. M., & Quinton, N. (1982). Pay and performance in major league baseball: The case of the first family of free agents. *The Journal of Human Resources*, *17*(3), 426-436.

Tukey, J. W. (1977). Exploratory data analysis.

Yang, Y., Yang, K. S., Hsann, Y. M., Lim, V., & Ong, B. C. (2010). The effect of comorbidity and age on hospital mortality and length of stay in patients with sepsis. *Journal of critical care*, *25*(3), 398-405.

Citation for this paper

"Outlier Identification in Count Data Using Variance Difference", Pelaez, A., Winston, E., Nejati, N., Zhu, X. 2018 North East Decision Sciences Institute, Providence, RI (April 2018)