The Definitive Pitcher Expected K% Formula

Mike Podhorzer

There have been many attempts to develop an expected pitcher strikeout percentage (xK%) formula, usually involving one of my favorite metrics SwStk%, perhaps average fastball velocity, and maybe another statistic or two. All of the regression equations did a fairly decent job, but there were always outliers, and I was beginning to see a commonality between them.

A couple of years ago, StatCorner included a pitcher's called and foul strike rate. For some reason, those metrics are no longer displayed on the player pages. However, when they were, they seemed to explain a lot of the discrepancies between the old xK% and actual K%. For example, Hiroki Kuroda and Jaime Garcia would consistently appear to be unlucky given their high SwStk% marks, but when digging further, you learn that they had gotten fewer called strikes than the average pitcher.

After this data disappeared, I vowed to find a new source. That source was eventually found, as Baseball-Reference displays the trio of "L/Str" (Looking Strike Rate), "S/Str" (Swinging Strike Rate) and "F/Str" (Foul Strike Rate) under the "More Stats" tab in the "Pitch Summary — Pitching" section. Initially, I was unable to find a leaderboard for these metrics, making it impossible to develop a regression equation. But then I realized that I should go directly to the founder himself. So I emailed Sean Forman and asked if a leaderboard existed for these stats, and sure enough, they do! The data isn't nearly as easy to manipulate as the exported stats from FanGraphs, but after much

effort, my data set was clean and ready for analysis.

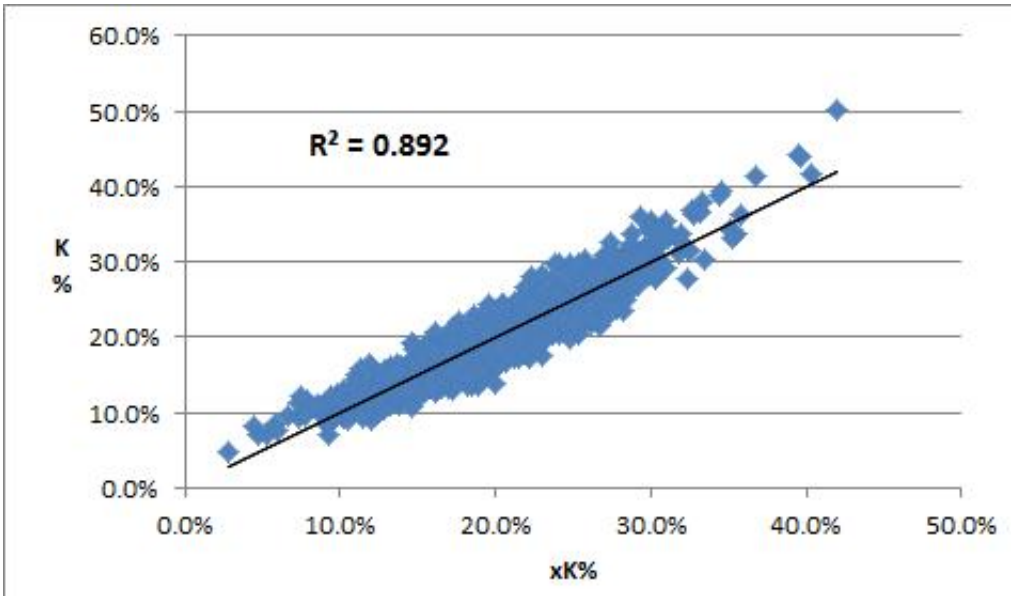
But first, why these metrics? Well, to record a strikeout, a pitcher needs to throw at least three pitches that result in a strike. Generating each of the three types of possible strikes are the underlying skills that directly lead to the ultimate result of the at-bat, the strikeout. While velocity is certainly important when predicting strikeouts, technically that should just lead to a higher swinging strike rate. So instead of including various metrics that merely predict the direct underlying metrics that feed into strikeout rate, it would be best to just use the direct underlying metrics to begin with.

I initially began my data set with all pitchers (starters and relievers, as there was no simple way to filter for only starters, and besides, I am not sure that would have even been necessary) from 2008-2012. While I would have liked to include as many seasons as possible, the work involved in cleaning each year just made it too time consuming. I felt like five years was good enough. I then narrowed down the data set to only those pitchers who threw at least 50 innings in a season. This left me with 1,629 pitcher seasons to analyze. Before we get to the results, the following table details the correlations between the three metrics and a pitcher's K%.

L/Str	S/Str	F/Str
0.01	0.81	0.20

It shouldn't surprise anyone that swinging strike percentage has such a high correlation. Heck, we could have only used that metric to predict K% and it would yield pretty decent results. I am, however, surprised that looking strike rate has essentially no correlation. This is especially true compared to foul strike rate, because a hitter obviously cannot strike out on a foul ball, but he could on a called strike.

Now let's get to the regression formula and graph.



xK% = -0.61 + (L/Str * 1.1538) + (S/Str * 1.4696) + (F/Str * 0.9417)

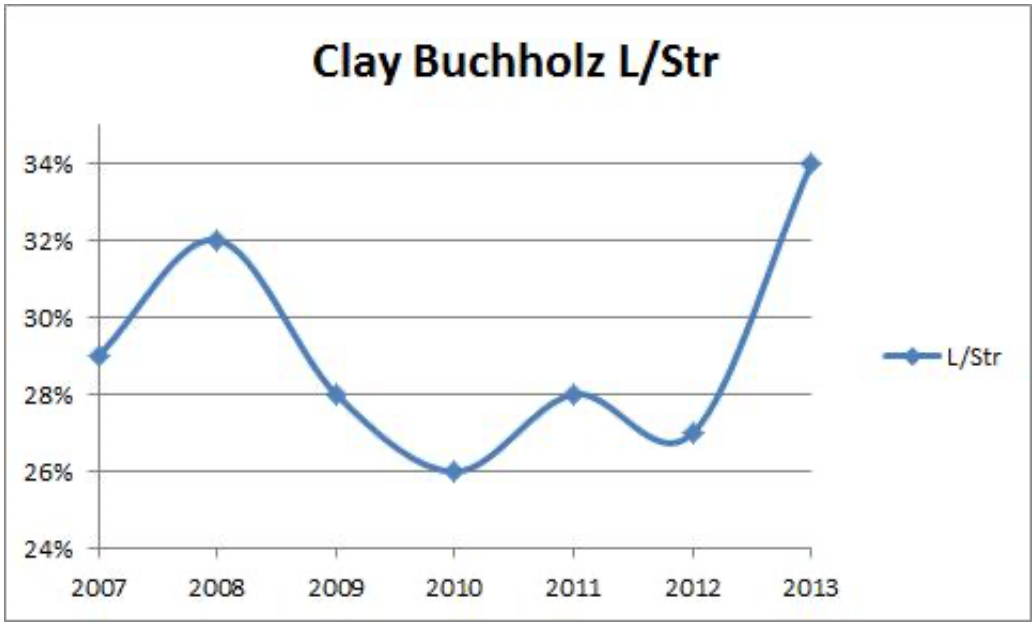
No, that was not a misprint on the graph. This equation, using the data I described above, produced an R-squared of 0.892. I would suggest that the remainder of the variation is just due to strike sequencing, which is essentially the luck component.

Interestingly, despite an essentially zero correlation, L/Str has a higher coefficient than F/Str in the regression. This makes much more sense as, once again, a pitcher cannot record a strikeout by inducing a foul ball.

In addition to developing an xK% formula, I was also interested in determining how repeatable these skills are. I was fairly confident that inducing swinging strikes was highly repeatable, but wasn't so sure about the other two strike types. So once again, I went back to my data set and narrowed the group to only those pitchers who had pitched in consecutive seasons, with at least 50 innings in both. That left me with n = 886. The following table represents the year-to-year correlations of each strike type rate.

L/Str	S/Str	F/Str
0.64	0.73	0.57

Again, it's no surprise that S/Str reigns supreme, but I did not expect the other two to rate so highly. If anything, these numbers suggest that pitchers are relatively consistent from year to year and do possess a high degree of control over these rates. So if you find a pitcher whose L/Str has suddenly spiked over a small sample of innings, it might not necessarily be such a fluke, but perhaps a true skills surge. Clay Buchholz is a perfect example of this scenario. Check out his L/Str rates over his career:



This season is the clear outlier, which may normally be chalked up to a small sample fluke. But given that L/Str rates do have a reasonable degree of consistency, it's very possible that Buchholz has taken this skill up a notch.

Next week, I'll put the 2013 data to work and report on the pitchers whose strikeout percentages are due for a surge or decline after consulting their respective xK% marks.