A clutch test, in brief

Paul T. Scott

September 4, 2008

I use a simple but powerful method to test the existence of clutch hitting in baseball and find *no evidence of individual clutch hitting*.[1] This brief paper is intended for those familiar with the clutch hitting debate and the tools associated with it. While this subject has been studied to death, I hope to offer a new and straightforward way of testing for the existence of clutch hitting.

There are two issues that must be addressed at the beginning of any study of clutch performance:

1. How to measure a player's performance, and

2. How to measure a clutch situation.

It is tempting and common to use a binomial statistic such as batting average (for 1) and divide plate appearances into "clutch" and "non-clutch" (for 2). Since the earliest clutch tests, both the definition of the clutch situation and the statistics available to evaluate a player's performance have evolved considerably. In older studies, the "close and late" definition of a clutch situation was often used, but it makes mistakes. For example, many close-and-late situations aren't as important as some close-and-early ones. Recently, a huge leap forward came in the form of Tangotiger's leverage index (LI), which provides a more-or-less continuous measure of the clutchness of a situation. Fortunately, LI removes the need for categorizing situation. Unfortunately, many studies continue the practice of categorizing plate appearances - perhaps out of habit, perhaps because it's easy to calculate the variance of a binomial statistic.

There is a sensible reason to categorize: some of the most popular measures of performance in baseball are percentages (e.g., batting average and on-base percentage). To see how these statistics change based on the clutchness of a situation, one must either group plate appearances according to clutchness or do a probit or logit regression with a continuous measure of clutchness like LI.[2]

---

[1] The data set includes 1997 and 2003-2007 data for players with at least 800 at bats. 1998 data was not available. Many thanks are due to retrosheet.org for providing the data.

[2] See Alan Jordan's study for the logit regression method: `http://www.tangotiger.net/Alan/ajordan.pdf`. His study, like mine, fails to find clutch hitting in Tango's data.

While I think the probit/logit method has potential, my method instead employs a simple linear regression with a single "clutchness" parameter for each player. One shouldn't do such a regression with a success/failure stat like on-base percentage; a more-or-less continuous measure of performance is more appropriate. A natural candidate here is *win probability added* (WPA), which measures how a player affects his team's probability of winning the game.[3]

However, raw WPA won't do, for it tends to zero exactly when LI tends to zero, no matter how good a player is. This is a problem because, if we simply look at WPA for different levels of LI, any good hitter will appear to be a very clutch one. For instance, if a player always hits home runs, his WPA outcomes from low-LI plate appearances will be very small (when a team is down 20 runs, even a home run does little to help the team's chances of winning). In a close game, the WPA outcomes from his home runs will be huge, so a naïve look at the player's data will suggest he is a very clutch hitter even though he just does the same thing all the time. A quick and easy fix for this is to divide WPA by LI, a statistic I call "adjusted win probability added" (aWPA).

My test involves a regression for each player:

$$aWPA_{i,j} = \beta_i + \gamma_i \cdot LI_{i,j} + \varepsilon_{i,j},$$

where β_i is interpreted to be player i's overall ability (measured in average aWPA per plate appearance) and γ_i is the player's clutch ability. A large positive γ_i indicates that batter i is a clutch hitter; a choker has a negative γ_i. $LI_{i,j}$ is the leverage index of player i's j-th plate appearance and $\varepsilon_{i,j}$ is the associated error term, which includes the contributions of the pitcher, the fielders, and randomness.

The null hypothesis that there is no individual clutch hitting in baseball is simply the hypothesis that all the γ's are equal to zero:

$$\mathrm{H}_0 : \ \forall i : \ \gamma_i = 0.$$

This isn't totally right, though. If better pitchers are pitching in more important situations, then we should expect hitters to be chokers on average. It turns out that there is a statistically significant "population choke" effect. That is, hitters in general tend to perform slightly worse in clutch situations than they do normally, probably because of the prevalence of strong relief pitchers and closers in clutch situations (managers seem to be doing at least one thing right). More formally, this is an endogeneity problem: the error term is correlated with the LI variable. This is easily corrected. I let $\bar{\gamma}$ represent this "population choke" or "Mariano" effect, and redo the regressions with the effect removed:[4]

---

[3] Typically (and in my study), a the probability of a team's winning is based on the half-inning, number of outs, runners on base, and score difference. The raw WPA of a plate appearance is simply the win probability after the plate appearance minus the win probability before the plate appearance.

[4] I actually run two large regressions: one with individual overall ability for each player and the population clutch effect, and the other with the individual clutch effects added. Then, I construct the F-statistic in the standard way: `http://en.wikipedia.org/wiki/F_test`.

$$aWPA_{i,j} = \beta_i + \gamma_i \cdot LI_{i,j} + \bar{\gamma} \cdot LI_{i,j} + \varepsilon_{i,j}.$$

Ultimately, I end up with an F-statistic, which can be translated into a p-value. Using the 1997 and 1999-2007 data (restricted to batters with at least 800 plate appearances), I find a p-value of .2122.[5] How unimpressive. I also applied this test to the same data set used by Tangotiger [3], finding a p-value of .6670,[6] which, unlike Tango's results, does not suggest that clutch hitting exists.

Thus, in contrast to the results by Tango [3], Dolphin [1], and in *The Book* [2], I find *no compelling evidence* for the existence of individual clutch hitting ability.

A virtue of this method is that WPA accounts for the full impact of a player's contribution. That is, in some situations, a sacrifice fly is just as effective at winning the game as a home run. In these situations, some measures of a player's performance may actually penalize the player for "coming through in the clutch" when he hits a sac fly rather than a home run. Any statistic like on-base percentage, OPS, or even runs created ignores the fact that the impact of a baseball event (in terms of WPA) depends on the situation. This method avoids this pitfall, for win probabilities will properly capture the impact of events like sacrifice plays. Those that believe clutch hitting has to do with adjusting to the situation at hand should find this method appealing.

On the other hand, this method also incorporates some extra variation, for occurrences that affect the win probability but are out of the batter's control (e.g., whether the runner on second is fast enough to make it home on a single) will still affect WPA. Whether this extra variation is worth the gains from capturing clutch performance and the clutchness of a situation more completely, I cannot say. Of course, one can always worry that this test lacks the statistical power to recognize the existence of a level of clutch hitting that we might consider important.[7]

References

[1] Dolphin, Andrew. "Clutch Hitting: Fact or Fiction?" http://www.dolphinsim.com/ratings/notes/clutch.html

---

[5] In this data set, there are 644 players, resulting in 645 parameters in the restricted model (model with no clutch coefficients) and 1288 parameters in the unrestricted model. The residual sums of squares are 3068.6206 and 3067.3492, respectively, resulting in a F-statistic of 1.0442.

[6] Tango uses 1999-2002 data, also restricted to players with at least 800 PAs. It contains 536 players, resulting in 537 parameters in the restricted model (model with no clutch coefficients) and 1072 parameters in the unrestricted model. The residual sums of squares are 967.5617 and 966.5693, respectively, resulting in a F-statistic of .9726. It's worth noting that Tango imposed the extra restriction that batters must have at least 100 clutch PAs. Since I don't categorize plate appearances in this way, I do not impose such a restriction, but it appears my data set actually contains significantly more batters.

[7] See "Underestimating the Fog" by Bill James, `http://www.sabr.org/cmsfiles/underestimating.pdf`.

[2] Tango, Tom M. , Mitchel G. Lichtman and Andrew E. Dolphin. *The Book: Playing the Percentages in Baseball.* TMA Press, 2006. Available at http://www.insidethebook.com/

[3] Tangotiger, "Does Clutch Hitting Exist? Yes!" http://www.tangotiger.net/clutch.html