- Search
  Search

- Fanshots
- Fanposts

- Best of Outnumbered

Shooting percentage regression

By Eric T. @BSH_EricT on Jun 13 2013, 4:23p 25

Richard Wolowicz

Over 125 years ago, it was demonstrated that small samples produce extreme results that will eventually regress to the mean. Is that true for hockey?

Tweet Share on Twitter (53) Share Share on Facebook (8) Share Share with Flyers friends 25 Comments ⋆ Rec Recommend this Post 3

The first paragraph of the wikipedia page on regression to the mean states, "to avoid making wrong inferences, regression toward the mean must be considered when designing scientific experiments and interpreting data."

Regression is a concept that is applied in all kinds of data sets involving complex human interactions -- medicine, sociology, economics, psychology, basketball -- yet for some reason, some people find it hard to assume it would apply to hockey. Hockey Wilderness considered the term a big joke last year, and now I find David Johnson of HockeyAnalysis.com doubting it in the comments of this article and a handful of tweets.

I assume they're not alone. So let's take a minute to discuss what regression means and why it is important to account for it.

Suppose I have all of you run a timed 40-yard dash right now. Who would have the ten worst times? It wouldn't be strictly the ten slowest people; there might be one person who slipped and two people who wore high heels today and one person whose breakfast burrito isn't sitting well and one person whose hangnail is killing them. On average, that group of ten that had the extreme worst performance likely got some bad breaks that put them down at the very bottom. If we have that group all run again next week, some might do even worse, but on average we'd expect the group to do slightly better.

So if we were trying to project their future performance, we would take their actual time and pull it up towards the average time a little bit, to account for the fact that someone who ended up down at the extreme was more likely to have been unlucky than lucky on that particular day.

How much of an adjustment would we make? That depends on how much luck we think there is. If we think there's almost no luck involved in a 40-yard dash time, then we'd make tiny adjustments. If we think there's a ton of luck, that'd lead to huge adjustments.

We can estimate how much luck there is by having *everyone* run the race again next week, and looking at how strong of a correlation there is between runners' first time and second time. And then there's some simple arithmetic that tells us how much regression we should expect: if the correlation between the results of the two races is r, then we should expect any given time to regress ($1-r$) toward the mean. So if the correlation is high -- let's say 0.9 -- then we might only expect movement 10% of the way back to the mean, but if it is low, those adjustments would be much larger.

So now instead of runners, let's talk about hockey players. You have probably all heard me say that shooting percentages are variable, so let's look at how that plays out.

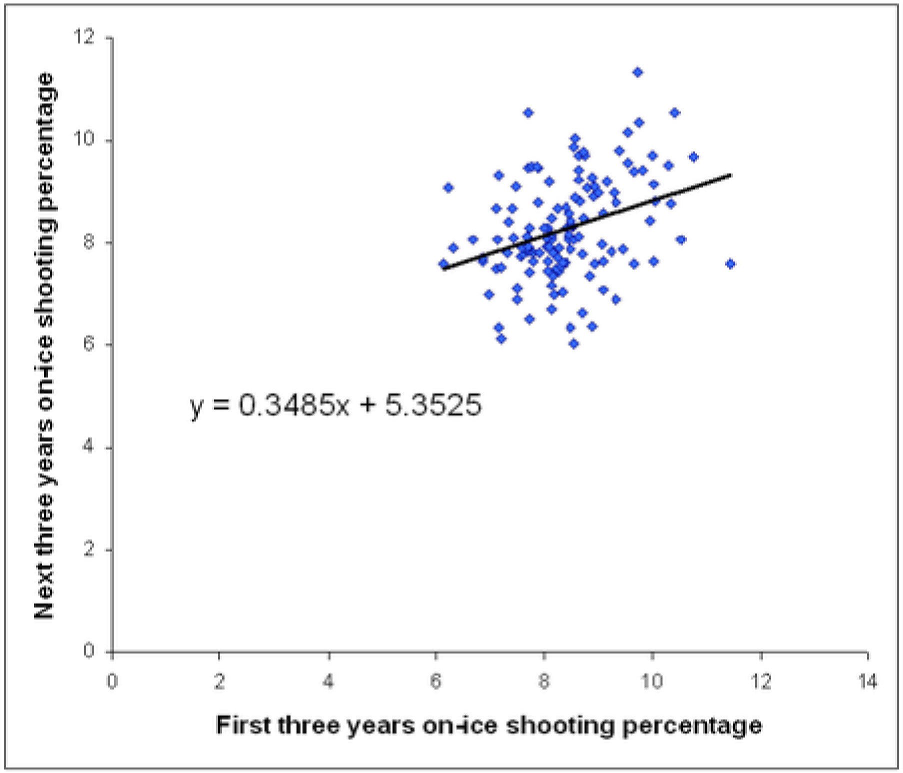
Let's suppose we take everyone who played at least 2500 minutes over the three-year period from '07-08 to '09-10 and also played at least 2500 minutes over the three-year period from '10-11 to '13. That's 129 guys, and we have what, at first glance, might seem to be a pretty large sample size. Each group averaged over 3000 minutes of ice time, which might seem like plenty for figuring out what the team's shooting percentage will be with them on the ice in the long run.

But here's the catch: 3000 minutes is a long time, but it represents only 120 or so goals, and so the unavoidable randomness of a lucky shot or a weird bounce or a goalie who briefly lost sight of the puck will have a sizable impact. It turns out that the correlation between their on-ice shooting percentage over the first three years and the next three years is only 0.33.

So our simple statistical rules tell us that if all we know is a guy's on-ice shooting percentage over a ~3000-minute sample, our best guess for what he'd do in the future would be to take that number and pull it back 67% of the way towards the average shooting percentage. We have no direct evidence that any particular individual got lucky or unlucky, so we'll be moving some guys in the wrong direction, but with this much luck involved, we'll win out in the end by assuming that everyone who's above average got at least a little bit lucky.

But that wasn't enough for David Johnson. He wanted me to prove that the rules of regression apply to hockey, just like they do to every other statistical endeavor. I feel like the burden of proof should be on him to prove that well-established rules don't apply rather than on the people using them, but since I have the data already prepared, I might as well post it.

Here's the scatterplot showing the relationship between on-ice shooting percentage over the first three years and on-ice shooting percentage over the next three years, complete with the line of best fit:



You'll notice that the line doesn't have a slope of one; it's much flatter than that. If you take a guy with an 11% on-ice shooting percentage in the first three years, the best-fit estimation of his next three years isn't 11%; it's (0.3485 * 11 + 5.3525) = 9.2%. Coincidentally enough, this is almost exactly 67% closer to the mean than the starting value was.

So there is no groundbreaking news here. The laws of arithmetic apply to hockey. In any measurement where the repeatability is not extremely high, predictions can be improved by regressing to the mean -- that's pretty much what a low repeatability means, after all.