Factoring regression into analysis

By Eric T. @BSH_EricT on Jun 16 2013, 4:58p 86

Jana Chytilova/Freestyle Photo

When several leading analysts make the same suggestion, you should at least look into it.

In the field of sports analytics, there has been a historical divide between the academic and online communities.

The online community often criticizes the academics for being behind the times, for using their high-powered tools to reinvent things that have long been known in Internet circles. I noted an example of that in an open letter published on Wednesday, discussing some adjustment the authors could make to have a greater, broader impact.

The academic community's response is that the lack of a formal peer review system makes the online efforts unreliable and making blogs

unacceptable as a serious research outlet. Tom Tango argued that this is unfair, that the informal peer review system online can be every bit as rigorous as the formal academic system.

My instinct was to agree with him, but this week I became aware of at least one case where our informal system is failing.

For over three years, David Johnson has been arguing that shot quality is an important factor. He has made some fine contributions along the way, but in at least one area he is following a path that no reasonable peer review system should allow. Rightly or wrongly, my credibility depends in part on the credibility of the online community as a whole, so I feel compelled to attempt to right this wrong.

On Thursday, I wrote about the importance of accounting for regression to the mean. This is a simple, fundamental element of analytical research. It is something they teach in the second month of a high school statistics class. It is not something that any peer review system -- formal or informal -- should allow an analyst to ignore.

I choose the word "ignore" carefully. This was not mere oversight.

As far back as 2010, such preeminent authorities as Hawerchuk and Tom Awad were cautioning Johnson that failing to regress his data to account for randomness was leading to dramatically overstated conclusions.

Even more direct than that early example is this one from two years ago (in the comments section):

> Hawerchuk: Keep in mind that *observed* performance is not equal to talent. You need to regress to the mean.
>
> David Johnson: Maybe I am an idiot and don't understand the regress to mean concept but if some guy posts an 11% shooting percentage year in and year out it seems to make sense to me that 11% is probably pretty close to his talent level.

I understand this is a reasonably subtle concept for someone who is not familiar with it, and it's not something I expect everyone to understand. But it's part of the price of admission if you want to be taken seriously as an analyst.

This has been a recurring theme with Johnson's work. Personally, the earliest I can remember cautioning him about it was when he pointed to Luke Schenn's reasonable goals-for rate and called him a "solid offensive contributor". I pointed out that this was entirely driven by a high on-ice shooting percentage that was not particularly likely to reflect any particular skill that Schenn might possess. If peer review were working for us, he would have investigated this at some point over the next eleven months.

Yet here we are today, with Johnson still steadfastly demonstrating a lack of understanding of the concept and insisting that "Regression to mean is worthless if everyone's 'true talent' is different" and "Shooting percentage has an uncertainty associated with it, regressing it doesn't necessarily make it better."

The whole point of having dialogue in comments and on Twitter is to get new ideas and strengthen your analysis. I can't even fathom simply

ignoring several different respected analysts who were telling me that a specific tool would markedly improve my work. Why engage in discussion at all if you intend to dismiss out of hand everything anyone else says, to insist that they show proof of basic statistical principles before you will even consider it?

I chose to humor him one last time, and demonstrate that on-ice shooting percentage does indeed regress exactly as expected.

His response?

> I am not convinced. My gut tells me that hockey is more complex and that a straight regression to the mean as you propose won't necessarily result in a better result.

Informal peer review is failing us. People read his work and take it seriously, not knowing the blatant disregard for statistical rigor that underpins it.

I write this article with two aims: 1) to shame Johnson for ignoring the suggestions of the community, in the hopes of changing his behavior and thereby reviving our ability to claim that our informal peer review is adequate, and 2) to show how the analysis ought to be done, in the hopes of ensuring that aspiring analysts have a model to follow and do not make this same mistake.

Let us move to phase 2: looking at how to improve a prediction of goal rate by accounting for shooting percentage regression.

In the previous article, we walked through how to calculate repeatability and thereby estimate how much regression should be expected. However, we can do even better with a more finely resolved estimate of repeatability, one that acknowledges that we expect more regression from players with smaller sample sizes and that accounts for a player's position.

Let's start with defensemen. There were 241 defensemen who played at least 50 5v5 minutes in both the 2007-10 and 2010-13 periods. We can sort them by minutes played in the first three-year period and divide them into five bins of about 48 players apiece. Then we can calculate how repeatable the on-ice shooting percentage was for the players in each bin and produce an estimate of the dependence on sample size. The table below shows the stats when the players in each bin were on the ice.

Bin	'07-10 shots	'10-13 shots	'07-10 sh%	'10-13 sh%	Repeatability
First 1/5th	179	665	8.1%	7.5%	**<0**
Second 1/5th	714	833	7.7%	7.6%	**0.10**
Third 1/5th	1213	979	7.8%	7.8%	**<0**
Fourth 1/5th	1559	1181	8.1%	7.8%	**0.00**
Last 1/5th	1922	1433	8.1%	7.8%	**0.09**

There is a lot of information there, but the key column is the last one: regardless of sample size, defenseman on-ice shooting percentage is essentially completely irreproducible. An estimate of how the team will perform with him on the ice in the future should include regressing the team's shooting percentage all the way back to league average -- which is essentially equivalent to using shot rates instead of goal rates.

And for the sake of people who are somehow unwilling to assume that shooting percentage regression will impact predictions of goal rates, let's compare using either past shot rate or past goal rate for predicting defensemen's future goal rate.

The table below shows the results of this comparison as a function of playing time. It's a little complicated, so I'll walk through what it means. Each cell represents a different threshold for how much ice time a player needs to have to be considered, and the contents of the cell compare the predictive power of shot rate and goal rate for that group of players. For example, if we include players with at least 1000 minutes in the first time period and at least 50 in the second time period, we see shot rate having a predictive power of 0.34 while that of goal rate is only 0.22.

Predictive power of shots / goals		Minutes played in '10-13		
		50+	1000+	3000+
Minutes played in '07-10	50+	0.28 / 0.12	0.36 / 0.16	0.41 / 0.28
	1000+	0.34 / 0.22	0.35 / 0.19	0.39 / 0.35
	3000+	0.40 / 0.30	0.38 / 0.31	0.45 / 0.40

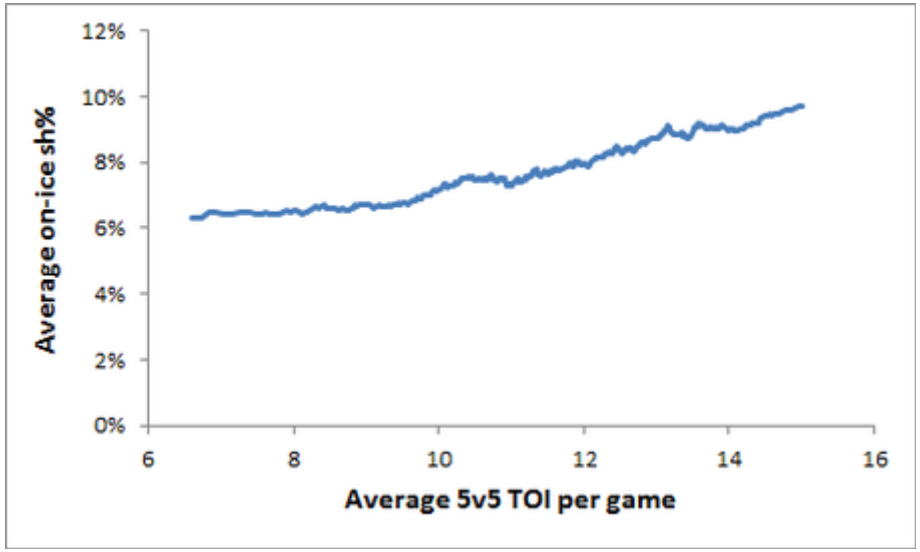
So save percentage is essentially wholly irreproducible for defensemen, and failing to account for that weakens the analysis dramatically regardless of sample size. Making predictions that assume Luke Schenn is somehow responsible for the team happening to have a high shooting percentage when he was on the ice in his first four seasons is tantamount to declaring a coin that came up heads four times in a row to be a magic coin.

For forwards, the story is different, of course. The range of shooting percentage talent at forward is much larger, and so persistent differences between players are observed. Still, those differences between players are significantly exaggerated by the role of variance.

Before we try to work out the repeatability for forwards, we should consider whether we might have cause to regress different players towards different means.

Tom Awad showed that there is a relationship between ice time and shooting percentage for forwards; a good shooter is likely to play with other good shooters, and their line is likely to get more than average ice time. So we should be able to do better than just regressing every forward towards the overall average shooting percentage.

Here is a plot showing shooting percentage as a function of average ice time per game:



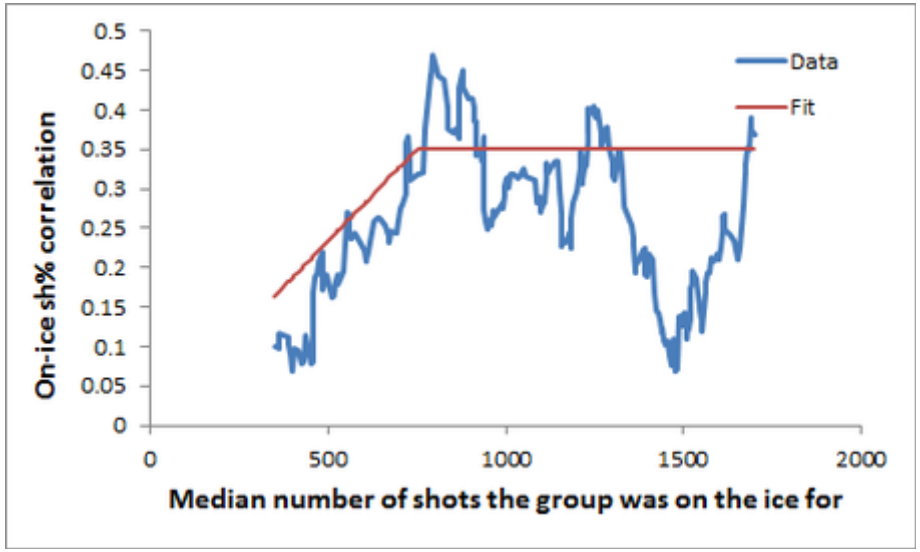
This plot gives us a starting estimate for each forward's shooting percentage. Before we know a thing about his shooting, we would estimate that a forward who averages 15 minutes of 5v5 ice time per game would have a ~9.7% on-ice shooting percentage, and that his observed performance in a non-infinite sample should be regressed towards that value.

This probably still isn't perfect. It is reasonable to guess that a player who runs hot for a while will get more ice time than his talent warrants, and a player who runs cold will get less ice time. The result then would be that the players who get 15 minutes of ice time likely have less than a 9.7% talent for shooting percentage in the long run, and we could do better by regressing their estimate a bit further. But this estimate will serve for a conservative first pass.

Now let's work on looking at repeatability as a function of sample size, so we know how far to regress each player towards the mean for players with their TOI.

Since we're no longer looking to regress everyone towards the same mean, we need to transform the data a little to accomplish this. Instead of comparing their shooting percentage in the two three-year time periods, we will compare their distance above/below their TOI-based expectation in the two samples.

Here is a plot of the repeatability of this metric as a function of how many shots a player was on the ice for in the first time period:



This plot is pretty noisy, which is unsurprising -- we're looking at low correlations, so without a ton of data, there's a lot of room for random chance to play a big role in any given point on this plot. I've arbitrarily picked a fairly simple function as a fit to this data -- again making a conservative estimate of how heavily to regress the data, leaving room for further improvements.

So now we can put the two plots together and make our projections. For any given player, we start with his on-ice shooting percentage in the first three years. We use his TOI to read from the first plot what the mean is for players like him, and we use his sample size to read from the second plot how far to regress towards that mean. That gives us an estimate of his future shooting percentage, which we multiply by his shot rate to project his future goal rate.

This is a conservative estimate of the potential for regression. We can refine the estimates of both his mean and the repeatability, and we could account for regression of his shot rate (which would be smaller because the sample size is much larger, but still isn't zero). We could separate out his own individual shooting percentage (which he has more control over) from his linemate shooting percentage (which regresses more heavily) and produce a better estimate of his future shooting percentage talent. But even with this conservative estimate, we see that applying regression has made appreciable improvements to the predictions:

Predictive power of regressed goals / goals		Minutes played in '10-13		
		50+	1000+	2000+
Minutes played	50+	0.62 / 0.55	0.66 / 0.58	0.56 / 0.48
	1000+	0.66 / 0.62	0.66 / 0.61	0.57 / 0.51

in '07-10	2000+	0.64 / 0.60	0.65 / 0.59	0.58 / 0.52

Simply making conservative estimates for shooting percentage regression substantially improves our projected goal rate, and dramatically alters our approach to evaluating defensemen. Frankly, I'm not sure how any other result could have been expected once we established that shooting percentage regresses heavily; the alternative would seem to require having a strong inverse correlation between shooting percentage and shot rate. It may be possible to account for some of the variance by examining usage (which might also be possible with the regressed version), but there necessarily must be a random component that should be accounted for.

Hopefully now we can put this three-year debate to rest and account for regression to the mean in any analysis that includes factors with poor repeatability.

More from *Broad Street Hockey*

- Craig Berube waited over a decade to punch Jeremy Roenick in the face
- Berube needs to ride his best players
- The Flyers need to double down
- Flyers vs. Rangers: Game 3 loss was not about 'effort'
- Gary Bettman discusses NHL expansion, shootouts, concussions and lots more

Latest News

- Lean on your best players, Chief
- Laughton recalled: Will he play?
- Mason returns for Game 4
- _laude vs. Claude: A poem in two parts
- The Flyers need to double down
- Thursday Morning Fly By