Adjusting for Score Effects to Improve Our Predictions

By Eric T. @BSH_EricT on Jan 23 2012, 12:00p 40

The analytical community believes strongly in shot differential as a predictor of future success. Unfortunately, the situation is complicated by score effects -- teams tend to go into a defensive shell when leading, which allows the other team to outshoot them.

To combat this, we typically focus on the shot differential when the score is tied to eliminate score effects. It has been shown that if you want to predict a team's future winning percentage, you should look at their shot differential with the score tied rather than at their place in the standings or goal differential.

This is very counter-intuitive, that goals and wins aren't the best measure of whether a team will get more goals and wins. But the problem is one of sample size -- it seems like a season is a long time and random bounces should even out, but with only two or three goals per game, they really don't. Because there are ten times as many shots as goals, random fluctuations in shot differential get much more rapidly washed out and we get a better measure of a team's talent.

So if sample size is so important, doesn't it seem inefficient to throw away all of the results from when the score isn't tied? Can't we find a way to correct for the score effects instead?

Sample size even plays a role in our decision about which type of shot differential we employ. In almost every case, we can improve SOG differential by also including shots that missed the net, which gives us a statistic called Fenwick. And usually we do better still by also adding in blocked shots to get the Corsi number.

This illustrates just how important sample size is. Blocked shots usually don't represent scoring chances, and as a result Corsi actually has a weaker correlation to goals in the long run. Yet even though it isn't quite as closely related to scoring chances, after 40 games or fewer, the larger sample size makes Corsi a better predictor than Fenwick.

As a result, I tend to rely on Corsi for most of my analysis. But wouldn't it be preferable to instead find a way to increase the sample size on Fenwick so that we can get an adequate sample size with the statistic that has the better true correlation to goal scoring? Let's see if we can do that by correcting for score effects so that we can include the non-tied situations and more than double the sample size on our Fenwick statistics.

Score-Adjusted Fenwick

Thanks to the Fenwick tabulations at Behind the Net, we can see that over the last four years, the average Fenwick for a team that's behind by two goals is about 56%. So if a team gets 57% of the shots when they're trailing by 2, that's 1% better than average -- just like if they had 51% of the shots in tied situations. Similarly, the average for a team that's behind by one goal is about 53.9%, so a team that gets 52.9% when down by one is 1% worse than average, like a 49% Fenwick Tied.

It's easy enough to make those corrections and come up with a Score-Adjusted Fenwick total that uses 42.4 minutes of even strength play per game, instead of the 17.9 minutes that goes into Fenwick Tied or 28.4 minutes that goes into Fenwick Close (another attempt to reduce score effects by focusing on close games).

This kind of correction has been suggested before, but there hasn't been a rigorous look at how much impact it has on the predictive power -- does it actually make our predictions significantly better? Let's look at the previous three years and see how it looks in comparison to Fenwick Tied or Fenwick Close.

Testing the predictions

One test of whether we've actually reduced the noise is called split half reliability, where we compare the team's score in odd-numbered games to their score in even-numbered games. The more a stat has random fluctuations, the lower the correlation between the two half-seasons will be. For Fenwick Tied, the R^2 is 0.70, while for Score-Adjusted Fenwick it is 0.82, so this is an encouraging start -- it seems that the larger sample size does reduce variance and allow us to better assess a persistent team talent.

The real question is whether that more precise measurement is actually more useful, whether we are measuring an important talent. We can start by looking at the correlation between Fenwick and point total. The R^2 for Fenwick Tied is 0.40, meaning that about 40% of a team's point total can be explained with Fenwick Tied. For Score-Adjusted Fenwick, that number increases to 46% -- including all of the game states gives us a better sense of how the team did.

Yet where we really hope to see an advantage of the larger sample size is with in-season predictions, as we might hope to make more precise assessments of talent and make them earlier in the season. To test this, we can compare the Fenwick score after a certain number of games to their points in the remaining games. Here's what we find as the predictive validity when we do that comparison:

	After 20 games	After 30 games	After 41 games	After 60 games
Fenwick Tied	0.40	0.38	0.34	0.28

Fenwick Close	0.43	0.39	0.34	0.27
Score-Adjusted Fenwick	0.46	0.41	0.36	0.29

Early in the season, Score-Adjusted Fenwick does a substantially better job of predicting how many points a team will earn in the remainder of the season. Later in the season the difference gets small -- both because the sample size for Fenwick Tied gets large enough that much of the noise is washed out and because the randomness of the small number of games remaining becomes increasingly difficult for any measure to predict.

Conclusion

Our ability to make accurate predictions depends on two things: having a good measure of talent and having a large enough sample size for that talent to dwarf random fluctuations.

Fenwick is our best inherent measure of a team's ability to get more scoring chances than their opponents, but analysts typically limit their sample size to situations where score effects will have limited or no impact. Instead, we can make very simple adjustments and include all situations, which more than doubles our sample size.

This gives us a good sample size with a good metric, resulting in better predictions earlier in the season.

★ ★ ★

Statistical post-scripts

Ideally, we'd merge the corrected Fenwick scores based on how much time a team spent in each game state; the more often a team leads by 2, the larger the up-by-2 Fenwick sample size will be and the more weight it should receive. However, since the TOI numbers at Behind the Net are buggy, I've used the league average TOI instead. The average team spends 3.75 minutes per game down by 2 goals, 8.46 minutes down by 1, and 17.94 minutes tied, giving us the following formula for Score-Adjusted Fenwick:

Score-Adjusted Fenwick = [3.75 * (Fen_up_2 - 44%) + 8.46 * (Fen_up_1 - 46.1%) + 17.94 * (Fen_tied - 50%) + 8.46 * (Fen_down_1 - 53.9%) + 3.75 * (Fen_down_2 - 56%)] / 42.36 + 50%

★ ★ ★

It is worth noting that the predictive powers reported here are smaller than those reported previously. The previous analysis was non-chronological; the test of the predictions after 20 games would ask: looking at the shot differential in 20 randomly-selected games, how well can we guess how many points the team scored in 60 other randomly-selected games?

That approach inherently reports stronger correlations than this method does because it washes out the impact of trades, injuries, coaching changes,

and so forth. For a team that made a mid-season trade, the non-chronological approach would look at the shot differential in a 20-game sample that included 10 games with the old roster and 10 with the new to predict the results in a 60-game sample that was similarly divided.

With the chronological approach, we are looking at 20 games with the old roster and trying to predict a 62-game sample that will be predominantly with the new roster. Obviously, the correlations will be lower, but I feel this more directly reflects the question at hand.

★ ★ ★

The thing that surprised me most in this study came from a multivariable regression of year-end Fenwick versus year-end points: Fen_up_1 had the largest coefficient and by far the lowest p-value (0.002, versus 0.10 for Fen_tied and 0.23 for Fen_down_1). Removing the weakest contributor (Fen_down_2) and rerunning the regression closed the gap somewhat, but it was still substantial, with p = 0.004 for Fen_up_1, 0.05 for Fen_tied, and 0.12 for Fen_down_1.

Given that the largest changes in expected points come from transitioning between the tied and down-by-1 game states, I would have expected performance in those game states to correlate more strongly to the standings. Is there some reason performance when up by a goal would be a particularly good measure of a team's talent? Or is this just a random blip? I'd love to get input on this.

More from *Broad Street Hockey*

- Remembering Homer: His 10 worst moves as Flyers GM
- Holmgren's 10 best moves as Flyers GM
- The Ron Hextall Era begins, and he says all the right things
- Here are some Vincent Lecavalier trade rumors
- Will Kimmo Timonen take a pay cut?

Latest News

- Tuesday Morning Fly By
- Homer's 10 worst moves