When presented the project of predicting three-point shooting percentages for NBA players, there were a number of directions I could consider. From my general awareness of statistical projection systems for baseball, the basis of the model would be to use past historical data to estimate a player’s true skill level. However, there are additional factors that could influence a player’s percentages. For example, a player’s true skill level can evolve over time, and while the direction and extent of that change may vary significantly by the player, there could be some generic trend evident across basketball. In addition, a player’s shooting percentages can depend heavily on the intra-game context of his shots, such as the location of the shots (distance or location along the arc), how open he is, and whether the shots are off-the-dribble or catch-and-shoot. Furthermore, there may be subtle inter-game influences such as whether the games occur at home or on the road and how much travel and rest time the player has had. Of the many different variables that could in theory impact three-point shooting percentages, many of them are either themselves unknown or their average effects may be determined to be minimal. As a result, the goal of this project was to build the model foundation that can predict three-point shooting percentages on its own and that can be extended in the future to include additional variables.

Baseball projection systems are well documented, such as [here](http://www.fangraphs.com/library/principles/projections/), [here](http://triplesalley.wordpress.com/2010/12/22/marcel-and-forecasting-systems/), and [here](http://www.hardballtimes.com/oliver-smarter-than-your-average-monkey/), and they all follow the same basic premise. Past historical data is weighted to estimate players’ true skill levels, with more recent years weighted more heavily; this is to account for possible random year-to-year variation while also allowing for more recent seasons to be more predictive than earlier seasons because true skill level can change. Then, the numbers are regressed to a mean, because through the sheer quantity of players and statistical variance, there will be outliers, and those outliers are not necessarily fully predictive of the future. Finally, an aging curve is applied to account for the average trend that young players tend to improve while old players tend to decline, so a player’s estimated skill level at the end of the prior year is not necessarily his estimated skill level for the upcoming season. One caveat is that many of these projection systems don’t deal with data before/outside of the Major Leagues, and as a result, players with no Major League service time (such as rookies) are treated as average players (in other words, they’re regressed entirely towards the aforementioned mean). While the different systems treat each of these steps slightly differently, they all use this general strategy. Accordingly, I’ve attempted to adapt this to basketball.

Before explaining the steps in more depth, here are a few general notes about my approach. All historical data is from basketball-reference.com. The html is scraped directly by R using the XML package. All model error calculations, such as when calculating the weighting coefficients/aging curve, are weighted by 3FA, so the testing error of the same magnitude is more significant for the player with more attempts. Only data from the 1998 season (which refers to the 1997-1998 season) onwards is used because the shortened three-point line in the 1995, 1996, and 1997 seasons led to much higher percentages; there was a similar dip in the 2012 season likely due to the lockout season that led to more compressed schedules and fewer rest days, but that dip was less significant and so the season is still included.



STEP 1: Weight Historical Data

Different baseball projection systems use different weights for each season (and those weights may further depend on the statistic being measured). Marcel (one of the simplest ones) weights the past three seasons 5/4/3 for all statistics. On the other hand Oliver uses 5/4/3 for hitters and 5/3.5/2.5 for pitchers, while Bill James weights up to eight years of data (the exact methodology is proprietary). Because there are many takes on the ideal weights to use and because there is reason to expect the ideal weights for baseball statistics might not apply to basketball ones, I’ve attempted to calculate my own weights.

There may be more robust ways to determine the weights, but a fairly simple method was to test various sets of integer weights and see which ones led to the lowest error. Specifically, sets of two, three, four, and five integers between one and eight were tried for players who’ve already been in the league for two, three, four, and five years respectively. Different numbers of years were used, because there are smaller samples sizes for three point attempts in basketball than there are for plate appearances in baseball and I was potentially concerned about possible selection bias from including only players who’ve been in the NBA for five years. I decided to start with the two-year grouping with the lowest error, take the three-year grouping that included the same weights for the most recent two years as before, and so on. This led to the weights (8,7,6,5,4) for the years (t-1,t-2,t-3,t-4,t-5). In reality, there isn’t a significant difference between the sets of weights as long as the recent years aren’t weighted less than earlier years.

STEP 2: Regression to the Mean

The overall concept was fairly simple; given some prior information about the population to which a player belongs and the player’s own smaller sample of performances, one would decide how much to weight each piece of information based on how reliable each estimate is, and this would depend on the amount of the statistic’s variance that is random and on the size of the player’s sample. Namely, variance in pitchers’ BABIP tends to be more random than variance in pitchers’ strikeout rate, so similar samples and discrepancies in BABIP would be regressed more heavily towards the mean; similarly, a pitcher with 2000 innings of an outlier BABIP would be regressed less than a pitcher with only 200 innings of that same BABIP.

The approach to use, however, was a lot more convoluted. Former colleague of Sam Hinkie, Eli Witus ,wrote an informative blog [post](http://www.countthebasket.com/blog/2008/05/19/regression-to-the-mean/) on various concepts and methods, some empirical and some mathematical, one could use to calculate regression to the mean. I tried a few of the mathematical ones to calculating reliability, but I was unable to figure out how to manage the very different numbers of opportunities by players in the data (some players shoot 700 3PA a season while others shoot 5) without getting unreasonable results. As a result, I decided to use the estimate by [Darryl Blackport](http://nyloncalculus.com/2014/08/29/long-take-three-point-shooting-stabilize/) via the year-to-year correlation method; a future next step would be to try this method myself. The estimate he provided was a reliability of 0.7 for 750 3PA. From this number and the equation

r = opps/(opps + constant)

I was able to determine the constant to be 321 and then use that constant to determine each player’s reliability from his number of 3PA. Then, each player’s three-point percentage was regressed (1-r)% to the mean.

The mean that was being used was the positional average shooting percentages (not weighted by attempts, so the average of 30% on 10 attempts and 40% on 700 attempts is still 35%). This was calculated using basketball-reference.com’s assignments of position. Oftentimes, when a player changes teams midway through a season, he will be then play a different position according to the data, and for this calculation, he’s included in both positions. Later on, when regressing specific players to their positional means, players with multiple positions are assigned the position for which they played the most minutes.

Because the number of 3PA by centers and the number of centers who take threes are so small, these positional means are very inconsistent:



As a result, I’ve decided to take the averages of the past five seasons (weighted by the coefficients earlier).



Now, the means are more stable and likely more representative of the true positional population means.

STEP 1’: Re-Weight Historical Data

Given the regression to the mean adjustment added from Step 2, it makes sense conceptually that recent years might need to be weighted more as there is relatively less of a need to strip out single-year outliers and, consequently, more of a need to capture potential changes in skill level in recent years. Doing so now leads to the weights (5,4,3) for the years (t-1,t-2,t-3). For the sake of simplicity, I’ve somewhat subjectively decided to now use only three years of data, because adding the extra years of data and their corresponding weights does not improve the error very significantly.

STEP 3: Aging Curve

There have been a number of different studies on aging in baseball and these studies used a number of different methods. The method that made the most conceptual sense to me is the delta method used by [Mitchel Lichtman](http://www.hardballtimes.com/how-do-baseball-players-age-part-1/). This method takes all couplets of players who recorded opportunities (which, for this study, would be 3PA) in back-to-back years and records the difference in their rates in the two years. Then, all of these couplets are summed by the ages of the players and weighted by the harmonic mean of their opportunities. The biggest issue with this method is survivor bias, in that players who “unluckily” recorded poor rates in the first year (and would normally perform better the next year through the simple fact that their expected ability is higher than their actual performance in the first year) might get fewer opportunities in the second year. Theoretically, a player who shoots poorly from the three-point line in one year would be discouraged to shoot the next year, but, then again, that’s never stopped [Josh Smith](http://www.basketball-reference.com/players/s/smithjo03.html#totals::none). Jokes aside, I haven’t adjusted for survivor bias in my calculations so far, and this would be something I would want to include in the future.

Using the delta method from 1998-2014, the curve appears to be linear.



Using varying degrees of polynomial regression (weighted by the number of attempts), the polynomial of degree 1 has the highest adjusted r-squared. Using 10-fold cross-validation to test out-of-sample, degree 1 also exhibits the lowest testing error. As a result, a linear model was fitted:



For the sake of curiosity, here’s how one might expect the percentage of a 30% 17-year old to progress over his career:



STEP 4: Backtest Results

In summary, my model weights a player’s three-point shooting percentages in the prior three seasons using the weights (5,4,3) for the years (t-1,t-2,t-3) to estimate his true shooting percentage. Because three point shooting takes a lot of attempts to stabilize, it then regresses each player’s percentage to his positional mean by (1-r)% where r = 3PA/(3PA+321). Finally, to estimate his percentages in his new season, the model adds/subtracts age-related improvement/deterioration from the linear model of best fit. For players without any attempts in the prior three years (either because they didn’t shoot any threes or because they are rookies/were out of the league), the model assumes the positional mean.

I’ve backtested the model for the years 2001-2014 using data available at the time. (However, to be entirely accurate, the weights/aging curve were not recalculated every year with only data available then.) Then, I did the same with basketball-reference.com’s [Simple Projection System](http://www.basketball-reference.com/about/projections.html) (while adding in the same positional mean assumptions for players whom SPS didn’t forecast due to lack of prior NBA data). It compares favorably to SRS:



STEP 5: Run Model for 2015

To run the model for the 2014-2015 season, I used the list of players on the roster section of basketball-reference.com’s individual team pages (i.e. <http://www.basketball-reference.com/teams/ATL/2015.html>). Potential issues are that this doesn’t include players currently not on a roster and these roster charts get changed anytime the team rosters change. As a result, the list of players I used may not be reproducible using the R code I ran.

To determine each player’s position, I used the position basketball-reference.com listed in the same roster chart mentioned above. In the cases players changed teams in their most recent season and played multiple positions, I used the position for which they played the most minutes. For rookies listed under combo positions, I manually used the position ESPN has listed for them.

NEXT STEPS:

There are a number of potential next steps for this project. One would be to improve some of the techniques used, in particular, determining the ideal weights to use for each season and calculating my own reliability for three-point shooting percentage. A second one would be to add additional inputs. Similar to how some baseball projection systems incorporate additional factors such as ballpark factors, I might consider the impact of some of the variables mentioned earlier. Specifically, a player’s three-point shooting is also dependent on the other players on the court such as his teammates. There may be ways to use methods similar to how regularized adjusted plus-minus is calculated to determine how each player affects the three-point shooting of his teammates. This would be useful in the extreme cases in which a player changes teams and is a proxy for the quality of the shots taken (off-the-dribble vs. catch-and-shoot and how contested the shot is). One such example would be Kevin Love between the 2014 and 2015 seasons; one would expect that the three-point shots he took in Minnesota (many of them created by himself) would be more difficult than the ones he’ll take in Cleveland (many of them off spot-ups next to LeBron James). If this approach proved impractical, another possibility would be to use expected usage rate as a variable (which would involve projecting usage rate), as usage rate could be another possible proxy for the quality of shots taken.