Three Point Percentage Prediction for the 76ers

Dan Weitzenfeld, dweitzenfeld@gmail.com

1 Executive Summary

A player's three point percentage is primarily driven by his skill and by the difficulty of the shots he takes. Because some players shoot more 'easy' shots than other players, we can't use three point percentage as a measure of shooting ability. To understand which players are better shooters, and to make accurate forecasts of shooting performance in the future, we need to disentangle skill and 'degree of difficulty.'

I developed a bayesian hierarchical model demonstrating the possibility of estimating a player's true, unobserved shooting skill, controlling for the 'degree of difficulty' of the player's attempts. The model shows us which players are over- and under-rated by three point statistics. When paired with predictions about the types of attempts a player will take in a given season - e.g. 50% uncontested corner, 25% uncontested arc, 25% contested arc - the model can be used to predict his observed three point percentage.

2 Data

I scraped data from basketball-reference.com. Specifically, I scraped the 'Totals', 'Shooting', and 'Play-by-Play' tables on the player-specific pages[1], for every player who played in the NBA since 2001. My dataset was therefore on a player-season level. I used python to perform the scraping, and I stored the scraped data in a local mysql database. For simplicity, I limited the dataset to players with at least 200 lifetime three point attempts.

3 Model Intuition

Three point percentages are determined by four factors: Skill, Degree of Difficulty, Injuries, and Noise. The ideal model will account for all four.

3.1 Skill

We can think of this as an unobserved 'true shooting skill.' Imagine if every player shot 10000 threes from the same set of places on an empty court - that measurement would be

[1] e.g.: `http://www.basketball-reference.com/players/c/curryst01.html`

the closest we could come to directly measuring shooting skill. Note that skill may change slightly over a player's career, but that once controlling for the other three factors, we wouldn't expect it to change significantly.

3.2 Situational Factors, or Degree of Difficulty

Of course, players don't shoot threes in a vacuum, and the situations in which they shoot vary significantly in terms of 'degree of difficulty.' Contested shots are more difficult than uncontested, arc more difficult than corner, pull-up more difficult than catch-and-shoot. The situations in which players make three point attempts are going to impact their observed three point percentage.

3.3 Injuries

An injury can cause a temporary or permanent drop in three point percentage. Think Kevin Love shooting poorly in Fall 2012, in between hand fractures. If we don't include the injury in our model, our estimates of Skill will be pulled downward for players who suffered an injury.

3.4 Noise

Whether we think about this as luck, or as capturing factors we can't possible measure, noise also contributes to observed shooting percentage. A model shouldn't be swayed too far by an outlier of a season.

Note that we observe Situational Factors and Injuries directly, but we do not observe Skill directly. But if we had an estimate of a player's Skill, an estimate of the impact of Situational Factors on shooting performance, and a prediction of the Situational Factors a given player is going to face in the upcoming season, we could predict his shooting percentage. We'd be able to make predictions like, "Player X's three point percentage is going to drop this year. In his career so far, he's benefitted from shooting uncontested corner threes, inflating his three point percentages: in reality, he's actually only a slightly-above average shooter. On his new team, he is unlikely to get so many open looks from the corner, so his three point percentage will fall to 30%."

4 Data Limitations and Situational Factors

Unfortunately, the only data on Situational Factors I could find was corner vs. arc attempts. But the model is easily expanded to include as many different types of shots (with different degrees of difficulty) as the data support. With access to SportVU data, for example, I would use 8 different shot types: contested/uncontested by arc/corner by catch-and-shoot/pull-up. **When you see arc vs. corner in the model, keep in mind that it's a small step from here to a more nuanced and accurate model.**

5 Model Implementation

When a model includes unobserved or 'latent' variables, bayesian inference is the most applicable technique. Bayesian inference has additional advantages, including quantifying uncertainty about model parameters, as I demonstrate in section 6, 'Results.'

I indicate the number of three pointers made from the arc by shooter i in season j as $y_{arc,i,j}$. These are modeled as independent Binomial:

$$y_{arc,i,j} \mid \theta_{arc,i,j} \sim \text{Binomial}(\text{attempts}_{arc,i,j}, \theta_{arc,i,j})$$

The corresponding variable, but from the corner, is $y_{corner,i,j}$. $\theta_{arc,i,j}$ and $\theta_{corner,i,j}$ are modeled:

$$\theta_{arc,i,j} = \frac{1}{1 + \exp(-1 * (\text{skill}_i + \text{injured}_{i,j} * \text{injury}))}$$

$$\theta_{corner,i,j} = \frac{1}{1 + \exp(-1 * (\text{skill}_i + \text{injured}_{i,j} * \text{injury} + \text{corner}))}$$

where:

- skill_i represents player i's unobserved Skill. Note that it doesn't vary season-by-season; this is a simplification, addressed in more detail in section 8.2, 'Age and Skill Drift.'

- $\text{injured}_{i,j}$ is 1 if player i was injured in season j, otherwise 0. It's observed, but because I don't have a dataset on injuries, I use whether the player played fewer than 30 games in the season as a proxy. This is weakness, addressed in more detail in section 8.3.

- injury represents the negative impact of an injury on a player's observed shooting performance. Note that this is a global variable: I've constrained every injury to have the same penalty. This is a simplification, addressed in more detail in section 8.3.

- corner represents the positive impact of shooting from the corner on a player's observed shooting performance. Note that this is a global variable: I've constrained every player to get the same benefit from corner threes. This is a simplification, addressed in more detail in section 8.1.

- the use of the logistic function ensures that the thetas are bounded $[0, 1]$

Player-specific skill levels are modeled as coming from a common distribution. (I used the python library pymc to perform inference, and I'm following pymc's convention of specifying the normal distribution as Normal(μ, τ), where τ is the precision of the distribution, defined as $\tau = \frac{1}{\sigma^2}$).

$$\text{skill}_i \sim \text{Normal}(\mu_{skill}, \tau_{skill})$$

The hyperparameters for the skill distribution are also modeled using non-informative priors. In layman's terms: I assume very little about player's different skill levels a priori.

$$\mu_{skill} \sim \text{Normal}(0, .001)$$

$$\tau_{skill} \sim \text{Uniform}(0, 100)$$

The remaining parameters are also modeled using non-informative priors:

$$\text{injury} \sim \text{Normal}(0, .0001)$$

$$\text{corner} \sim \text{Normal}(0, .0001)$$

A graphical representation of the model is depicted in Figure 1.

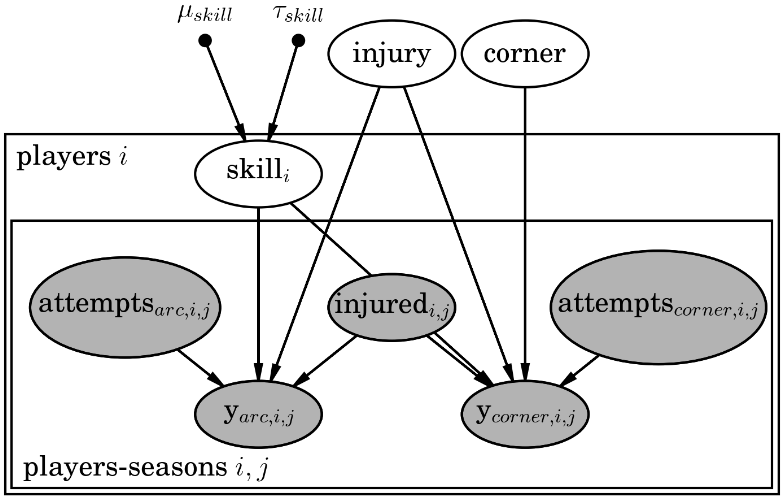

Figure 1: The DAG representation of the model. Injury and corner are global, skill is at the player level, and all observed variables are at the player-season level. Grey shading indicates observed variables, while no shading indicates the latent variables we are inferring from the data.

6 Results

First, let's look at the variable 'corner,' to determine the posterior likelihood of the benefit a shooter receives from shooting corner threes over arc threes (Figure 2).

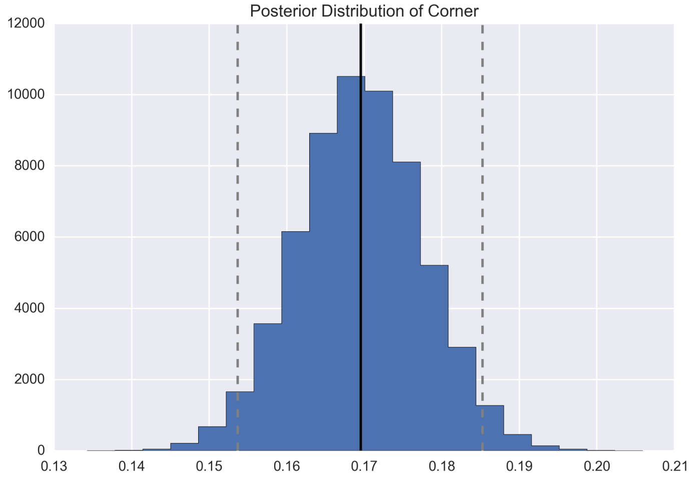

Figure 2: Posterior distribution of the variable 'corner.' 95% HPD interval implies a 3-4 percentage point increase in three point percentage if a player went from shooting 100% arc to 100% corner attempts.

As expected, the posterior distribution on corner implies that shooting from the corner has a lower 'degree of difficulty' than shooting from the arc.

Because some players shoot more 'easy' shots than other players, we can't use three point percentage as a measure of shooting ability. The model estimates true shooting ability, controlling for attempt degree of difficulty. Figure 3 visualizes this idea, by graphing lifetime three point percentage against the means of the player-level posterior skill distributions. Players denoted by a red dot benefitted over their career to date from taking attempts with a low degree of difficulty, while those denoted by green suffered from taking attempts with a high degree of difficulty.

Lastly, let's take a quick look at the injury variable - see Figure 4.

The posterior distribution on injury implies that shooting percentage in an aborted season is likely to suffer, but only a tiny amount. As discussed in section 8.3, I'm using aborted season as a proxy for injury, so these results should be taken with a grain of salt.

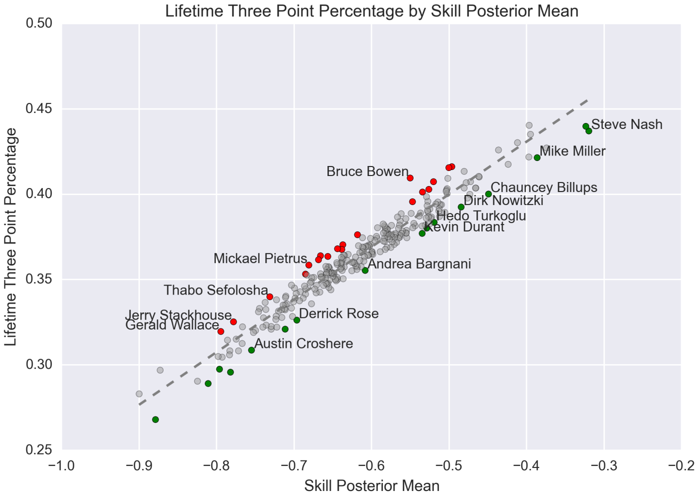

Figure 3: Plot of lifetime three point percentage, by player's mean posterior skill level. Players denoted by a red dot benefitted from taking attempts with a low degree of difficulty, so their observed three point percentage overstates their true shooting ability. Players denoted by a green dot suffered from taking attempts with a high degree of difficulty, so their observed three point percentage understates their true shooting ability.

7 Predictions

Note that my model makes explicit the impact of situational factors on a player's observed three point percentage. To make accurate predictions, we ought to predict how these situational factors will change. This is a strength of the technique: it enables us to quantify our uncertainty about the our predictions. For example, we might make an educated guess that Kevin Love will shoot between 5% and 15% of his attempts from the corner this season; the bayesian model allows us to easily integrate that belief into our predictions. For example, Figure 5 shows how our predictions of Alexey Shved's three point percentage in 2014 vary by our estimate of the proportion of his attempts he will take from the corner. Because predictions of 2014 situational factors are beyond the scope of the exercise, I'm not going to make predictions for every shooter, but the examples above demonstrate the technique and the advantages of making predictions in a bayesian framework.

8 Weaknesses and Future Work

If given an opportunity to work further on this model, I would address these weaknesses:

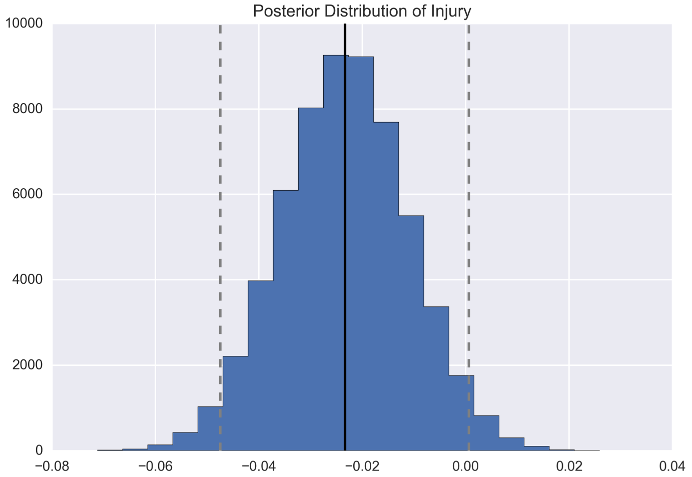

Figure 4: Posterior distribution of the variable ‘injury.’ 95% HPD interval implies a 0-1 percentage point drop in three point percentage if a player is shooting ‘injured’ (in an aborted season).

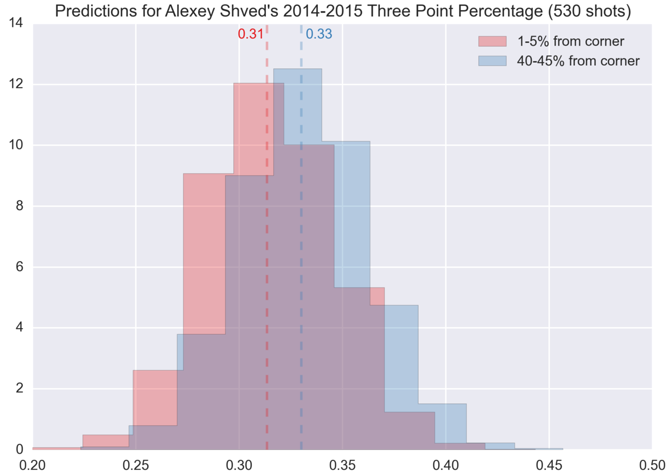

Figure 5: Using posterior distributions of the skill and corner variables, along with our belief about the degree of difficulty of his attempts, we can simulate Alexey Shved’s three point percentage for the 2014-2015.

8.1 Situational Factors, or Degree of Difficulty

As I described in section 4 ‘Data Limitations,’ the biggest limitation of the model is the simplified representation of the Situational Factors, a.k.a the Degree of Difficulty. With access to better data, I would be able to break down three point attempts into more categories, better capturing the varying degrees of difficulty.

Moreover, I’ve constrained the benefit of shooting from the corner to be at the highest level of the model hierarchy: I’m not allowing it to vary by shooter. I’d like to experiment with player-specific corner variables, to see if the evidence supports some players receiving a bigger benefit from shooting corner threes than others.

8.2 Age and Skill Drift

In this model, I’ve constrained a player’s skill level to be constant over their career. This is obviously an oversimplification, and the model could account for skill drift over time.

8.3 Injuries

Using a proxy (whether a player played more than 30 games) for injury is an obvious weakness. There’s even a danger that the proxy is actually picking up on situations in which a player is shooting poorly and is subsequently benched. I’d like to replace this proxy with actual data on periods when players were known to be suffering from injuries.

8.4 Rookies

The model does not make predictions for rookies. A different model - one that takes into account performance in college, and historical data on the college-to-NBA transition - would be required.

8.5 Degree of Difficulty is Exogenous

In the model, I treat shot choice (degree of difficulty) as if it’s imposed upon players from on high - as if the player himself has no say in his shot choice. This is true to the degree that the coach’s strategy controls the shots taken by a player; to the degree that players choose to shoot in a given situation, it’s a weakness in the current model. Future modeling - especially with access to SportVU data - could integrate a players’ shot/pass decisions, making degree of difficulty more endogenous to the model.

9 Code

My scraping and data-manipulation code can be found here: `https://github.com/DanielWeitzenfeld/three-seven-six`. My modeling and visualization code can be found here: `http://`

nbviewer.ipython.org/github/DanielWeitzenfeld/three-seven-six/blob/master/notebooks/model2-v2.html

Let me know if you have problems with either of these links.