Modeling College Basketball Shot Location Data

Ryan J. Parker
rjparker@ncsu.edu
ST 733

Technical Report
April 19, 2011

Abstract

In this study we consider how to model a college basketball team’s shot attempt success probabilities from various locations on a basketball court. This data has been traditionally collected by coaching staffs, but the data is now being made more available to the public through the internet. With this relatively new source of data, it can be seen that although the raw data is helpful, there are some areas for concern. For example, there are some locations where shot attempts have not been recorded. To alleviate this and other issues, this study illustrates how to spatially model this data with a CAR model to create a more accurate look at a team’s probability of making a shot from a given location on the court versus those models that ignore the spatial information in the data.
## Contents

1 Introduction .................................................. 3

2 Shot Location Data ........................................... 3

3 Modeling Shot Success ........................................ 4
   3.1 Sample Means by Location .................................. 4
   3.2 Probit Regression ........................................... 5
       3.2.1 Intercept Only Model .................................. 5
       3.2.2 Home Court Model ....................................... 6
       3.2.3 Model Comparison ....................................... 7

4 Summary .................................................... 8

References .................................................... 9

Figures ....................................................... 10
   Duke’s Shots .................................................. 10
   Angle/Distance Grid .......................................... 10
   Duke’s Raw Data in Angle/Distance Grid ..................... 11
   Intercept Only Model ........................................ 11
   Shot Success: at home ....................................... 12
   Shot Success: not at home ................................... 12
1 Introduction

In the game of basketball each team’s primary way of scoring points is by attempting shots from various locations on the basketball court, and publicly available data is starting to be collected that tracks the location of each of these shots. The goal of this study is to model this data to provide better estimates of team shot success from each location of the court. This information would be useful to a college basketball coaching staff, as smoothed estimates of shot success are likely to be more useful than the raw estimates that suggest a bumpy transition from location to location to go along with having some locations where the data estimates attempted shots will never (or always) be made.

Motivated by the work of Reich et al. (2006), we will consider probit regression models using the CAR distribution to spatially smooth regression coefficients. We will consider multiple models, and 5-fold cross-validation will be used to estimate model prediction error to help select a model for this data. Going into this study, we expect that a model that incorporates spatial information with regards to where shots are attempted will perform better than those models that ignore this spatial information.

2 Shot Location Data

From CBSsports.com (2011), shot location data from the 2010-11 season has been collected that contains over 96,000 shot attempts from 265 teams in 855 NCAA division I basketball games. Because the focus is on how a coaching staff might use this data, we will only consider the shot location data for Duke, the atlantic coast conference (ACC) team with the most observed shot attempts. This data can be downloaded from:

http://www4.ncsu.edu/~rjparker/duke_shots.csv

These over 2,100 shot attempts from 37 of Duke’s games are represented in Figure 1. In this graph, shot attempts are shown with respect to where they were attempted on the basketball court. Shot attempts represented by an “o” were successfully made, while those represented by an “x” were missed. This graph illustrates that, unfortunately, this data set contains errors. For example, the shots at the top of the graph suggest that shots were attempted out of bounds. Also, there are some shots that are attempted from a very far distance from the basket. Results of these kinds of shots are not likely to be interesting for a coach to consider, and are thus ignored in this analysis.
3 Modeling Shot Success

Intuition suggests that shot location data lends itself to a spatial analysis, as we would naturally expect shots further away from the basket to have a lower chance of success versus those closer to the basket. Also, we would expect that shots taken from similar locations on the court will have a similar probability of success.

To model shot success we define a spatial grid to combine shots into more manageable components for performing an areal data analysis. First, we consider a grid using angles and distances as defined by Reich et al. (2006). This 11x11 spatial grid, shown in Figure 2, also contains an approximately 3-feet by 3-feet region around the rim. In this grid we define neighbors to be cells that come from adjacent angles and distances. For example, cell 59 has distance neighbors 58 and 60, and it has angle neighbors 48 and 70. Note, however, that the region that contains the shots closest to the rim is defined to have zero neighbors. This is because these shot attempts are the least like those attempted from other regions, as they mostly come in the form of dunks and layups, shots that have a high probability of success versus those from other locations on the court.

3.1 Sample Means by Location

Figure 3 shows Duke’s sample means from each location on this grid. This graph illustrates the difficulty with using the sample means from each cell in the grid to represent success from that region of the court. One difficulty is that some regions indicate no shots were attempted from the region, thus there is no estimate for these cells. Another difficulty is that there are regions with estimates suggesting high probability of success that are next to regions with estimated low probability of success. Intuitively we would expect a smoother transition from one region to the next.

To make inference with this data we use a probit regression that makes use of the CAR distribution to spatially smooth regression coefficients. Within this framework we can make inference with regards to models of the data, and we can also estimate the prediction performance of these models.
3.2 Probit Regression

To model this data we construct the following probit regression model for \(n\) observations and \(p\) predictors that we fit using Markov chain Monte Carlo (MCMC) methods\(^1\) using 10,000 iterations and a burn-in of 1,000:

\[
\begin{align*}
\Pr(Y_i = 1) &= \Pr(Z_i > 0) \\
Z_i &\sim N(X_i \beta_{s(i)}, 1) \\
\beta_j &\sim \text{CAR}(\tau_j)
\end{align*}
\]

for \(i = 1, \ldots, n,\ j = 1, \ldots, p,\) and where \(s(i)\) is the region associated with observation \(i\). The \(\text{CAR}(\tau_j)\) prior distribution for the coefficient for predictor \(j\) in region \(k\) can be defined as:

\[
\beta_{jk} | \beta_{jl,l \neq j} \sim N(\bar{\beta}_{jk}, 1/([\tau_j m_j])
\]

where \(\bar{\beta}_{jk}\) is the mean of \(\beta_j\) at the \(m_j\) neighbors of region \(j\). Also, a \(\text{Gamma}(0.01, 0.01)\) prior is placed on \(\tau_j\).

3.2.1 Intercept Only Model

We first consider an intercept only model for this data. This model has \(p = 1\), and so using the form of Model (1) this model is specified by having \(X_1\) be a vector of 1s. Fitting this model, we have that the posterior median of \(\tau_1\) is 120.8, with 95% credible interval being (24, 384). For example, a region with 4 neighbors has an estimated smoothing standard deviation of 0.045, with a 95% credible interval for this smoothing standard deviation being (0.026, 0.102).

The results of this model fit are graphed in Figure 4 with estimated probabilities of shot success on the left and posterior standard deviations on the right. This graph illustrates that this model estimates that Duke is a team that shoots better from the right side of the court from the perimeter, as that side of the court contains numerous regions that have an estimated probability of shot success between 0.39 and 0.41 versus the numerous regions on the left side of the court that have estimated probability of shot success between 0.37 and 0.39 (including a few regions between 0.35 and 0.37).

This difference is practically significant, as some of those shots are behind the three point line meaning they are worth an extra point when successfully

---

\(^1\)Special thanks to Dr. Brian Reich for providing code to fit this type of probit regression model in R.
made. Thus a 3pt shot on the right side of the court with a 0.41 probability
of success is worth the same number of points as a 2pt shot with a 0.615
probability of success. Alternatively, a 3pt shot on the left side of the court
with a 0.37 probability of success is worth the same number of points as a 2pt
shot with a 0.555 probability of success. Hence the Duke coaching staff would
want to create 3pt shots shots on the right side of the court for their team
or, alternatively, determine and correct the reason for their lower shot success
from the left side the court.

The posterior standard deviations in Figure 4 for these estimated probabil-
ities of shot success show that the shots along the angle closest to the baseline
(those near the very top of the graph) have the highest amount of uncertainty,
where as regions near the middle of the court where more shots are taken have,
as one might expect, lower standard deviations.

3.2.2 Home Court Model

In college basketball teams tend to play better when they are at home. Thus
a coach may be interested in knowing how their team plays at home versus
when they do not play at home. So in this case we will consider a model using
as a predictor an indicator for when Duke is playing at home. Again, using
the form of Model (1), this home court model is specified by having $X_1$ be a
vector of 1s and $X_{i2}$ being equal to 1 when Duke’s $i^{th}$ shot attempt is at home
and 0 otherwise.

Fitting this model, we have that the posterior median of $\tau_1$ is 103.7 and
$\tau_2$ is 21.9. These have 95% credible intervals of (19.7, 431.6) and (4.9, 195.1),
respectively. For example, a region with 4 neighbors has an estimated smooth-
ing standard deviation for the intercept and home court parameters being 0.05
and 0.11, respectively.

The models associated with this fit are shown in Figure 5 (Duke at home)
and Figure 6 (Duke not at home). In each graph the estimated probabilities
of shot success are on the left and posterior standard deviations are on the
right. The first thing to note is the higher amount of uncertainty for shots
at home compared to those not at home. This is likely due to the combined
uncertainty associated with the parameter indicating when Duke is playing at
home with the uncertainty of the parameter for the probability of shot success
when Duke is not at home.

The estimated probabilities of shot success suggest that Duke shoots better
from 3pt range at home versus when they are not at home. Also, this model
estimates that when at home Duke does not shoot as well from mid-range, as
their at home graph shows a lot of dark and light blues for these areas versus
a majority of white for these areas when not at home. It’s not clear why this would be the case, as we would expect Duke to shoot better from most regions of the court when at home versus when not at home. Given the amount of uncertainty in these parameters, this unexpected difference is likely due to a confounding variable, such as quality of opponent.

### 3.2.3 Model Comparison

In this section we compare the models above with a couple of simpler models of the data. First, we fit a model that ignores location (NOLOC), where the estimated probability of shot success for every attempted shot is simply the sample mean of the shot results. Second, we fit a model that only differentiates between 2pt versus 3pt shots (2 vs 3). Again, this model uses the sample mean of the observed shot results for 2pt versus 3pt shots as the estimated probability of shot success for each respective shot.

To estimate the prediction error of these models we use 5-fold cross-validation. Error is assessed using $-2 \log [\Pr(\text{result})]$, where $\Pr(\text{result})$ is the probability assigned to the observed event “result” (Hastie et al., 2009). The table below shows total and mean cross-validation error for each of these models:

<table>
<thead>
<tr>
<th>Model</th>
<th>Total CV Error</th>
<th>Mean CV Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOLOC</td>
<td>2707</td>
<td>1.38</td>
</tr>
<tr>
<td>2 vs 3</td>
<td>2668</td>
<td>1.36</td>
</tr>
<tr>
<td>Intercept Only</td>
<td>2591</td>
<td>1.325</td>
</tr>
<tr>
<td>Home Court</td>
<td>2590</td>
<td>1.325</td>
</tr>
</tbody>
</table>

The first two models listed in this table are for the simpler models of the data. As we would expect, these simpler models are outperformed by the intercept only and home court models (although not shown, the estimated cross-validation standard errors confirm statistically significant differences). That said, there is no clear winner between the intercept only and home court models. Based on these results, we would not confidently prefer to use one model over the other. In fact, it would probably be illustrative for a coach to examine results from both models to get a sense for where their team may have an advantage they want to exploit or a weakness they want to correct.
4 Summary

In this study we examined inferences a CAR model provides for college basketball shot location data, and we illustrated that CAR models are better than models that assume no (or little) about the spatial structure of the data. Also, we were unable to find that a model that takes home court into account to be decidedly better than a model that does not. Ultimately, however, a coach would probably want to utilize both models when examining their own or an opposing team.

One area for future study would be looking into other important predictors of shot success, such as opposing team. Also, one might consider modeling shot success for each player on the team. The distribution of where shots are attempted is also important, a modeling problem that Reich et al. (2006) tackle using CAR models. Thus incorporating inferences between a shot success and shot distribution model is also an area for future study.
References


Figure 1: Duke’s Shot Location Data

Figure 2: Basketball court with angle/distance grid
Figure 3: Duke’s raw data in the angle/distance grid

Figure 4: Duke’s smoothed probability of success using the intercept only model (left) and associated posterior standard deviations (right)
Figure 5: Duke’s smoothed probability of success at home (left) and associated posterior standard deviations (right)

Figure 6: Duke’s smoothed probability of success when not at home (left) and associated posterior standard deviations (right)