Applying Bayesian Statistics in Predicting NBA Basketball Prospect Outcomes Using Aggregate Scouting Information

Ted Henson

8/8/2020

Motivation

This goal of this project is to apply Bayesian statistics in the rethinking package by Richard McElreath from his book Statistical Rethinking. There are many different techniques, packages, and software that use Bayesian statistics or techniques. This analysis will focus on Bayesian regression, quadratic approximation, causal inference, and more as I continue to develop and apply my understanding of Bayesian statistics. There are an infinite number of causes or variables that could potentially predict NBA success if one had access to the information: work ethic, receptiveness to coaching, wingspan, vertical leap, top speed, collegiate or international basketball success. Some of these variables such as collegiate success (measured by statistics) are available to the public. Some such as vertical leap may be available to a team after a workout or NBA combine. This analysis is not meant to create the best possible predictions for a prospect as that would require a lot of time and resources. The goal is to apply Bayesian techniques in a basketball setting and to see how much influence a player's high school recruiting ranking plays in NBA success.

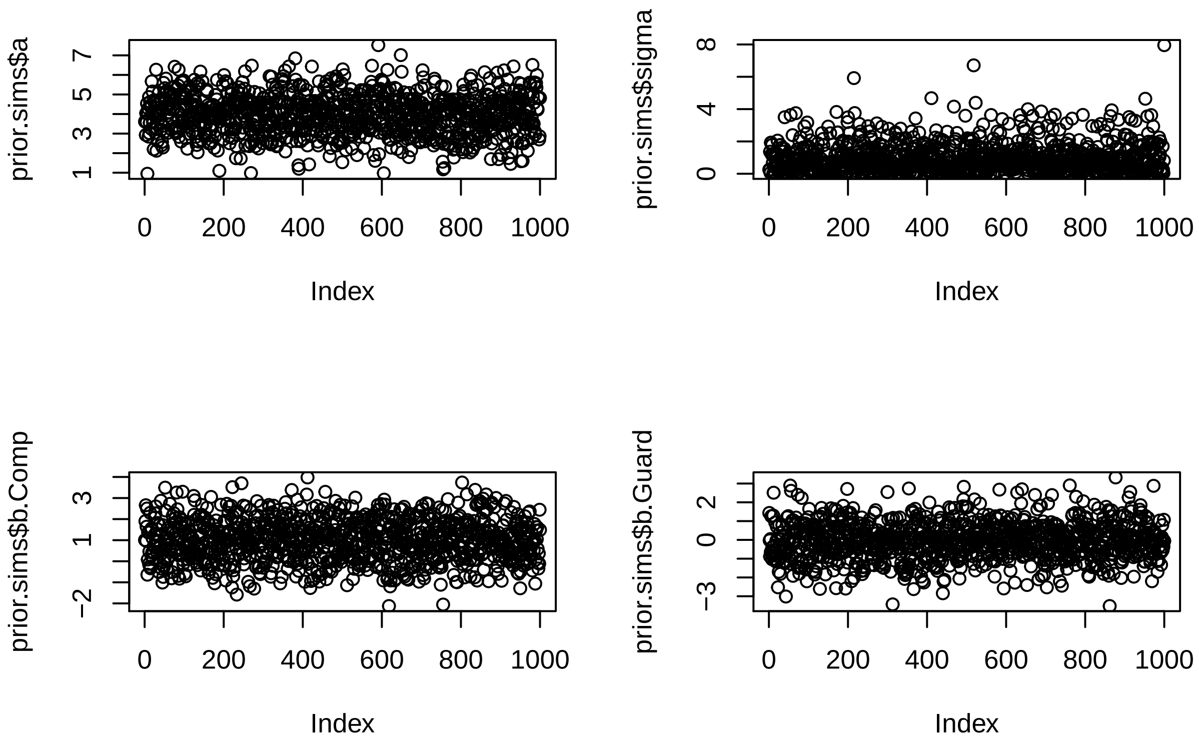
Data

The input data is aggregate scouting information from the Recruiting Services Consensus Index (RSCI). The data includes rankings from a variety of recruiting services for the top 100 high school prospects going back to 1998. Rivals is the only service that has been constant through the years. ESPN has been in the index since 2013, but their initial creator, Dave Telep, has been ranking prospects since 1998 so they are treated synonymous for this project. Other services that are still active include 247Sports and Hot100Hoops. All other services in the index no longer rank prospects. The variables used in the analysis include the composite ranking and the services still active. There little reason in running analysis on variables if they no longer exist. Also included in the the predictor matrix was a player's age and position. The response data is the advanced metrics table from basketball reference. This table includes statistics such as Value Over Replacement Player (VORP), Win Shares (WS), and many others. This analysis will attempt to first predict Win Shares Per 48 Minutes (WS/48) for a player's rookie season. Rookie seasons were chosen to simplify the analysis. WS/48 was chosen as opposed to VORP as VORP penalized rookies who played a lot of minutes on bad teams. In order to control for outliers who played very few minutes, only rookies who played over 100 minutes were considered. If a player did not play over 100 minutes they either were injured, were not given a chance, or were ineffective when they played. In practice one would not exclude them and either include them, regress their stats towards the mean or minimum, or something else, but for simplicity they are excluded here.

```r
source('~/Bayesian Statistics with RSCI Hoops/Code/RSCI to NBA Merger.R')
source('~/Bayesian Statistics with RSCI Hoops/Code/Bayesian Regression.R')
```

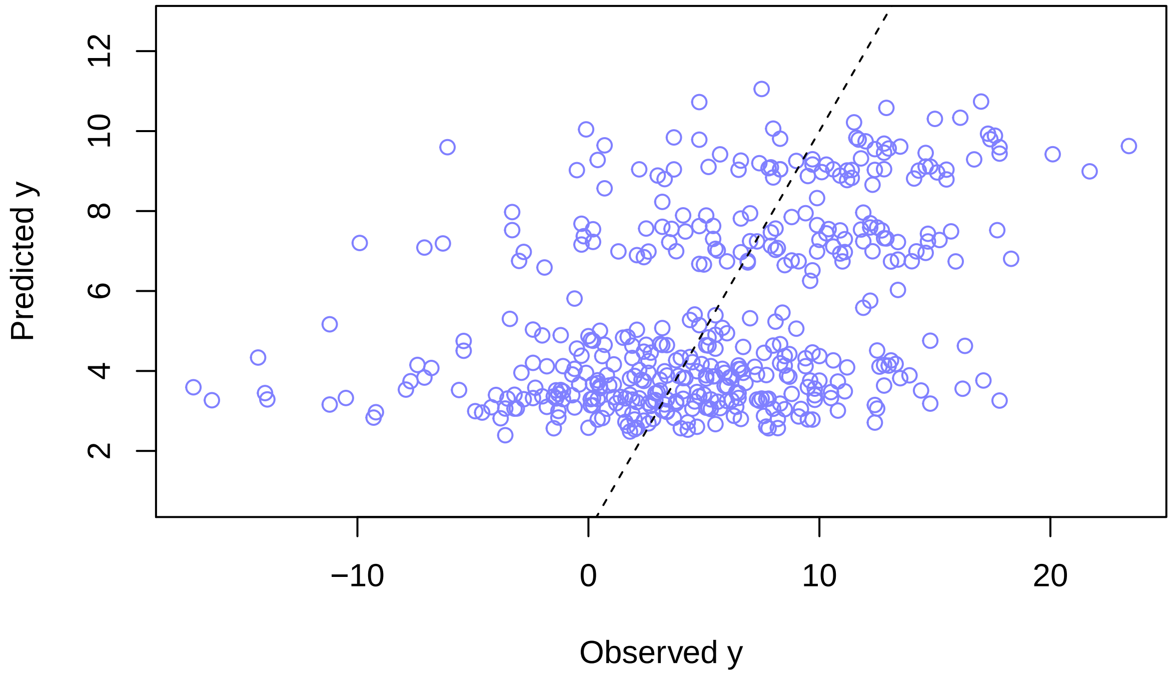
Plots of Simulation of Priors

```r
par(mfrow=c(2,2))
plot(prior.sims$a)
plot(prior.sims$sigma)
plot(prior.sims$b.Comp)
plot(prior.sims$b.Guard)
```



Plot of Observed versus Expected WS/48 with Predictive Intervals

```r
plot(mu_mean ~ basketball$y,
     col = rangi2,
     ylim = range(mu_PI),
     xlab = 'Observed y',
     ylab = 'Predicted y')
abline(a = 0, b = 1, lty = 2)
```


Predicted y
2
4
6
8
10
12
−10
0
10
20
Observed y