Investigating Effect of Recruiting on Team Success in NCAA Football

Nick Berry

May 8, 2015

1 Introduction

The college football season is an annual gridiron of 120+ teams putting it all on the line for a chance to be named the nation's best team. During the year each team plays 12 games and losing a single one can ruin your chances of winning the National Championship. While the actual season is unequivocally exciting, the off-season for college football fans also offers its share of intrigue.

College football teams don't have a yearly draft and can't offer their players money, so they have to figure out other ways to find quality players and entice them to come play for their school. This process is called recruiting. Whether it is the playing time, a national championship, or a good draft position in a few years, colleges and their coaches will promise almost anything to get the players they need to win. There are college scouts that sit through hundreds of high school games and practices a year to assign grades to players and report those grades back to their head coaches. All of this effort is put into recruiting because coaches know that the only way to win is to have great players.

We will look at how college football teams have performed on the recruiting trail, especially focusing on how individual past years' classes affect a current year's teams' skill. Along with that, we will compare the actual quality of a team with the expected quality based on the recruiting classes they have had.

2 Data

There are countless websites dedicated to grading recruits and rating entire classes for colleges all over the country. One such site is Rivals.com, a Yahoo! affiliate. Rivals publishes individual rankings for recruits on a star scale with a 1-star being a low ranked recruit and a 5-star being the best of the best. For perspective, there are about 30 5-star recruits every year and about 250 4-star players, making up $\approx 1\%$ and $\approx 1\%$, respectively. The individual player scores are then aggregated into a team score using a formula that is basically a linear combination of the star rankings of your recruits and a couple of special bonuses. We will use this team score as a metric for the quality of recruiting classes. Rating of college football recruits by outside entities has only been a practice since around 2002, so we have data from then on. In order to make the results more informative, the team class scores are normalized within each year, so a score of 0 is average.

We are looking to use the recruiting class quality to predict the actual quality of a team, so we need a numerical quantifier of a team's skill. Jeff Sagarin created a ranking system that is widely known and was even used by the NCAA to rank college football teams for a while. Sagarin's rankings give each team a score, usually in the range of about 50 to 95, that can be used to compare teams within a year of play and even predict score differential if teams were to play. We will treat Sagarin's rankings as the truth for how good a team was at the end of the season.

Although Sagarin has published rankings all the way back to 1998, we only use 2006 - Present because we need the 5 previous years of recruiting data to predict each year of responses.

3 The Model

We wish to create a model for how the 5 previous years of recruiting affect a teams current skill. This lends itself naturally to a linear regression situation where we get a coefficient for each year's contribution to the team's Sagarin score. Initially, we use a model with a fixed intercept and only look at the past class' effects. This model will evolve into a more flexible random intercept model that will also give us some team specific information. We will use a Normal data model, which is a justified choice based on Figure 1.

The point of this analysis is two-fold. First we want to examine only how recruiting classes effect a teams success. This can be viewed in its most pure form using the model with a fixed intercept. The second fold is the team specific aspect of the model. We want to analyze how different teams have fared relative to their expectation based on recruiting classes and which teams would be expected to perform the best if each team had equal recruiting classes.

3.1 Fixed Intercept Model

Let $i = \{1, \ldots, 122\}$ represent team and $j = \{2006, \ldots, 2014\}$ represent the year. Then,

$$Y_{i,j} \overset{\text{ind}}{\sim} \text{N}\left(\mu + \beta_1 x_{i,j} + \beta_2 x_{i,j-1} + \beta_3 x_{i,j-2} + \beta_4 x_{i,j-3} + \beta_5 x_{i,j-4}, \sigma^2\right)$$

where $Y_{i,j}$ is the Sagarin score for team i in year j, $x_{i,j-k}$ is the recruiting score for team i, k years before year j, $\boldsymbol{\beta}$ is a 5 dimensional vector.

We take μ, $\boldsymbol{\beta}$ and σ to be independent, so

$$\begin{aligned} \text{p}(\mu, \boldsymbol{\beta}, \sigma) &\propto p(\mu)p(\boldsymbol{\beta})p(\sigma) \\ \text{where } p(\mu) &\propto \text{N}(70, 5) \\ p(\boldsymbol{\beta}) &\propto \text{N}(\mathbf{0}, \tau^2 \mathbb{I}_5) \\ p(\sigma) &\propto \text{Cauchy}^+(0, 3) \end{aligned}$$

As a prior for the variance of the distribution of βs we will use the prior $p(\tau) \propto \text{Cauchy}^+(0, 1)$.

This model formulation results in estimation of parameters μ, $\boldsymbol{\beta}$, σ, and τ. μ is the intercept, which represents the mean overall Sagarin score. The parameters β_1 to β_5 are the contributions to the team's end of year Sagarin ranking for the first year to fifth year players on the team, respectively. σ is the standard deviation of the differences between the actual Sagarin scores and the scores predicted by μ and the βs.

3.2 Random Intercept Model

Like the fixed intercept model, we have a β vector that gives us slopes for each year of recruiting and a prior on that β vector. The random intercept model, however, assigns an intercept to each team rather than a single overall intercept. Again, we let $Y_{i,j}$ be the Sagarin score for team i in year j and $x_{i,j-k}$ be the recruiting score for team i, k years before year j. The model is:

$$Y_{i,j} \overset{\text{ind}}{\sim} \text{N}\left(\mu_i + \beta_1 x_{i,j} + \beta_2 x_{i,j-1} + \beta_3 x_{i,j-2} + \beta_4 x_{i,j-3} + \beta_5 x_{i,j-4}, \sigma^2\right)$$

Now, we assign the following independent priors

$$\begin{aligned} \mathrm{p}(\mu_i, \boldsymbol{\beta}, \sigma) &\propto \prod_{i=1}^{n} \left[p(\mu_i)\right] p(\boldsymbol{\beta}) p(\sigma) \\ \text{where } p(\mu_i) &\propto \mathrm{N}(\mu, \tau_2^2) \\ p(\boldsymbol{\beta}) &\propto \mathrm{N}(\mathbf{0}, \tau_1^2 \mathbb{I}_5) \\ p(\sigma) &\propto \text{Cauchy}^+(0, 3) \end{aligned}$$

In addition,

$$\begin{aligned} p(\mu) &\propto 1 \\ p(\tau_1) &\propto \text{Cauchy}^+(0, 1) \\ p(\tau_2) &\propto \text{Cauchy}^+(0, 1) \end{aligned}$$

4 Results

4.1 Fixed Intercept Model

For the fixed intercept model I ran 10000 iterations of MCMC with 1000 iterations of burn-in. The trace plots mixed well and all signs point to adequate convergence to the target distributions. The effective sample sizes are all very large and the $\hat{R}$ values are close to 1. We have no reason to suspect that the chain is inadequate or didn't converge.

The results of the chain are as follows:

	2.5%	50%	97.5%
μ	70.09	70.63	71.17
β_1	2.72	3.94	5.17
β_2	0.66	1.87	3.09
β_3	0.51	1.70	2.89
β_4	-0.02	1.11	2.25
β_5	-1.18	-0.05	1.04
σ	8.65	9.02	9.41
τ	1.30	2.21	4.66

Table 1: Table of 2.5^{th} percentile, the median, and the 97.5^{th} percentile for posterior distributions of parameters for the fixed intercept model.

The plot of the densities for the parameters $\beta_1 \ldots \beta_5$ can be found in Figure 2. The posteriors reveal a median of 70.6316567 for the common intercept. The plot of the βs shows that β_1 is significantly higher than any of the other βs, while β_2, β_3, β_4 are clumped as a middle group, and finally that β_5 is the lowest of the βs. The median posterior value of σ is 9.0185027.

4.2 Random Intercept Model

In order to get estimates for the random intercept model, we have to obtain posterior distribution samples for 132 different parameters. We accomplish this, as above, via Markov Chain Monte Carlo. I ran the model with 10000 iterations including 1000 burn-in iterations. As above, the trace plots mixed well, the effective samples sizes are large, and the $\hat{R}$ values are all very near to 1. There is no reason to suspect that the chain doesn't converge to the target distribution.

	2.5%	50%	97.5%
β_1	1.94	2.99	4.03
β_2	0.56	1.58	2.61
β_3	0.71	1.69	2.74
β_4	0.44	1.40	2.40
β_5	-0.55	0.40	1.34
σ	7.16	7.48	7.84
τ_1	1.13	1.90	3.95
τ_2	4.38	5.11	6.01
μ	69.56	70.58	71.57

Table 2: Table of 2.5^{th} percentile, the median, and the 97.5^{th} percentile for posterior distributions of parameters for the random intercept model.

Estimates for means of the distributions for μ_i can be found in Table 1 at the end of this report. Estimates for the non-μ_i values are in the following table.

The predicted random quantities for this model are similar to the fixed model. The βs look like the fixed model's β posteriors, having at least the same general shape. μ again is a representation of the overall mean, specifically it is the mean of the distribution of posterior means. The 122 posteriors for different μ_is represent distributions for individual schools' intercepts. The posterior median of σ is 7.4791547, which is about 1.5 points lower than the median estimate of σ for the fixed intercept model.

5 Discussion

The MCMC for both of the models behaved correctly, giving us no reason to think that our chain didn't converge or that there were any computational problems of any kind. Now we can start to actually use the results that we have spent this entire time trying to obtain.

5.1 Fixed Intercept Model

In the first, fixed mean, model we were primarily concerned with looking at how previous year's recruiting classes effect the current team's success. Looking at Figure 2, we see that β_1 is the largest slope by a significant margin. β_1 is the slope corresponding to first year players on the team. A large slope means that the first year players have a large impact on the quality of the team. In more detail, β_1 represents the predicted change in a teams Sagarin score for a 1 unit increase in the recruiting score for the most recent recruiting class, when all other predictors are held constant. The other βs have similar meanings, with the corresponding recruiting class being altered instead.

Since β_1 is significantly larger than the other βs, we can conclude that a teams success is dictated largely by its Freshman class. This makes sense from the perspective that Freshman come into a program raw and untrained, meaning that there is a large difference between the best Freshman and the worst Freshman. This may not be true as students get older. The practices for every team in this study are rigorous, the weight training is top of the line, and the coaches all know what they are doing. I believe that the skill levels of the players start to shrink together, making the effects of a class lessen as those students grow older. Of course there are still really good players, but those really good players don't always correspond to the really good recruiting classes. College athletes develop in different ways throughout their career, but Freshman are, for the most part, as advertised during recruiting, they haven't had a chance to change or develop yet.

The β_2 and β_3 distributions for slope are nearly identical with the distribution of β_4 being slightly

shifted to smaller values from β_2 and β_3. The slopes for all 3 of the middle βs are definitely positive, but are not as high as the β_1 value. This is good because it means that the recruiting classes for 2 years, 3 years, and 4 years prior do have a positive relationship with Sagarin rankings, and can give us information about expected quality of a team.

Finally, β_5 is the expected increase in Sagarin rankings corresponding to a 1 point increase of the recruiting class that are now 5^{th} year seniors, if all other covariates are held constant. The distribution of β_5 is centered around 0 and gives us no evidence that it has an effect on Sagarin rankings. I included the 5^{th} year coefficient in the model because a significant portion of college football players stay through their 5^{th} year and I wanted to be able to answer whether even after some of their class was gone, their recruiting class was still having an effect. It doesn't.

5.2 Random Intercept Model

Although not officially what I wanted to look at when I first started this model, I find that the interpretation for the intercepts are among the most interesting information revealed by this analysis. If each team was given an equal and average recruiting class, (each year's covariate equal to 0) the intercepts represent that teams expected Sagarin score. So it is basically a measure of how well a team has used its talent. We could look at the slopes like we did for the fixed intercept model as well, but if those quantities are of interest then we are better off just using the fixed intercept model to examine them.

It is most interesting to think of it as a "coaching effect," in the sense that the intercept quantifies how good my team is thanks to only the coach, independent of the players' skill levels. If a team has a very low intercept, it means that they had low Sagarin scores despite decent to good recruiting. This could reflect poor development of players or poor coaching and training during the season. If a team has a very high intercept, it likely comes from them getting poor to mediocre recruiting classes but performing well and getting high Sagarin rankings with those below average players. These schools get the best out of their players. In essence, a high intercept means that you performed better than expected based on your recruiting, and a low intercept means that you performed worse than expected based on your team's recruiting.

Looking at point estimates for the intercepts, a couple of interesting points arise. 5 of the top 10 schools as far as intercept are concerned (Boise State, BYU, Cincinnati, Utah, and Navy) are small schools that play in poor conferences. This is the perfect situation for those teams because they generally recruit at an average to below average rate because of the size of the school, but they play in bad conferences so they still win a lot of games and earn respectable Sagarin rankings at the end of the year. Among the 5 teams listed above, Navy is in a different situation than the others. Navy perenially ranks in the bottom 5 in recruiting classes and then performs at a just below average rate, but they are so bad at recruiting that even having a Sagarin ranking just below average is a feat for them. Boise State, BYU, Cincinnati, and Utah are all decent sized school that recruit at about an average (maybe just below) rate and perform very well, contending for well known bowl games and even being in talks for BCS bowl games in the past.

Other interesting intercepts are Alabama with the 33^{rd} highest. Alabama is always near the top in recruiting and is year after year a top 10 team in Sagarin ranking, but even Alabama's success couldn't get them ranked higher on the list because of how good their recruiting is. Kansas State is ranked 13^{th} on the list under Bill Snyder, who has baffled people for more than 20 years by turning seemingly talent starved teams into winning juggernauts. My alma mater, Texas A&M, is ranked 52^{nd} which is just about average. They get just above average recruiting classes and perform at a just above average level in general. My three favorite rankings are Texas at 97^{th}, Michigan at 111^{th}, and Miami (FL) at 112^{th}. All three are traditional powerhouses and get great recruiting classes every year, but have really been struggling for the past 6-7 years. They perform much worse than they should with the talent that they bring in, and it shows in their intercepts.

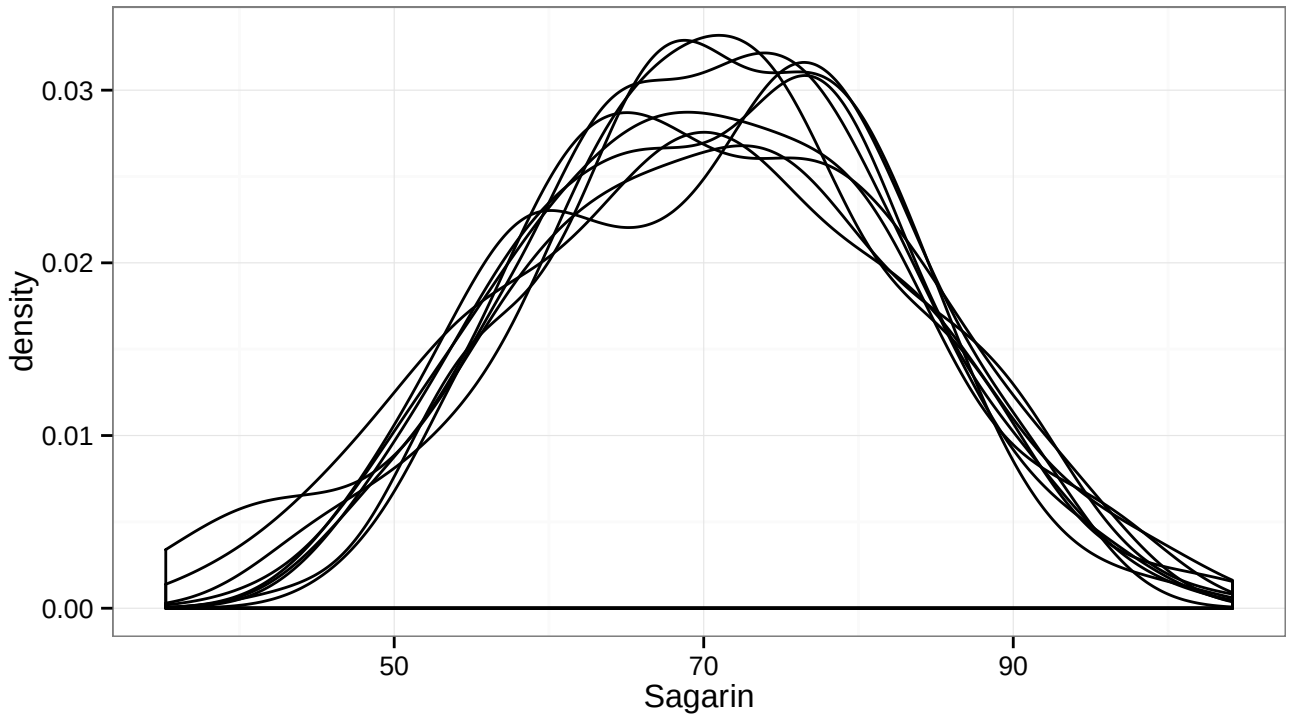

Figure 1: Densities for the distribution of Sagarin's rankings for NCAA football teams. Each curve represents a different year. Verifies that a normal data model is appropriate for modelling Sagarin rankings.

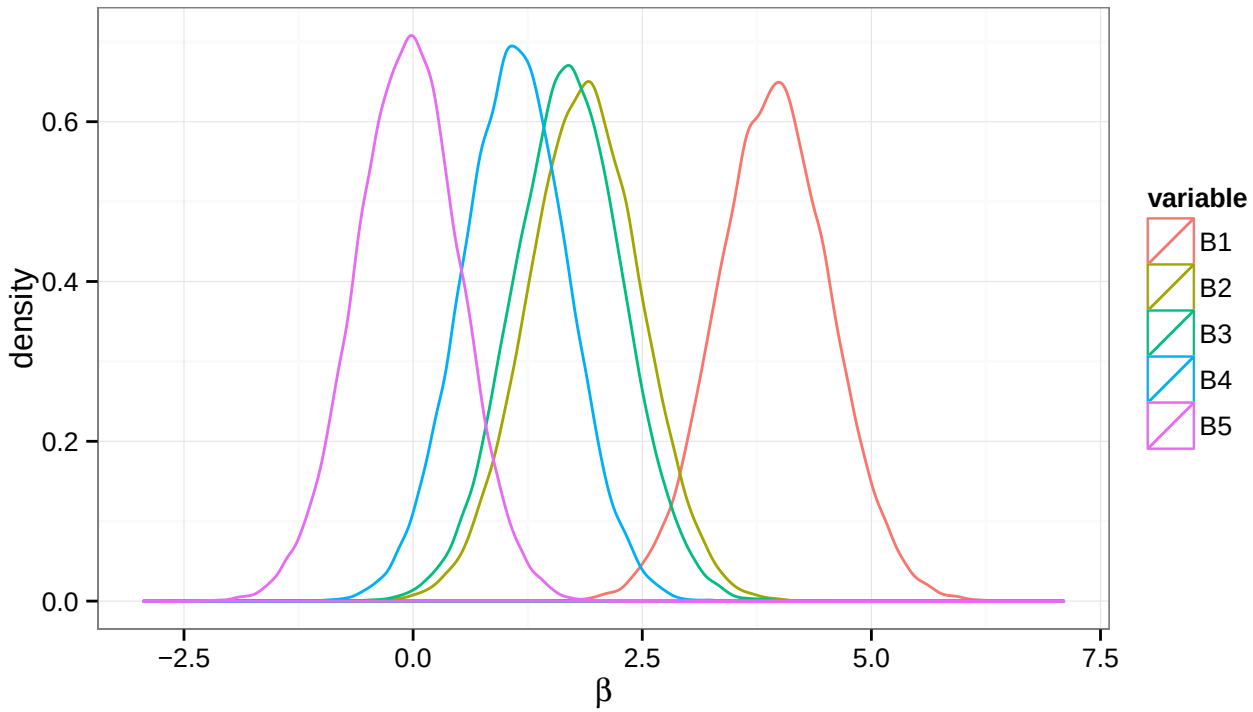

Figure 2: Densities for the posterior distributions of $\beta_1 \ldots \beta_5$ for the fixed intercept model. Notice the 3 sets of distributions: β_1 to the far right, $\beta_2, \beta_3, \beta_4$ in the middle set, and β_5 centered around 0.

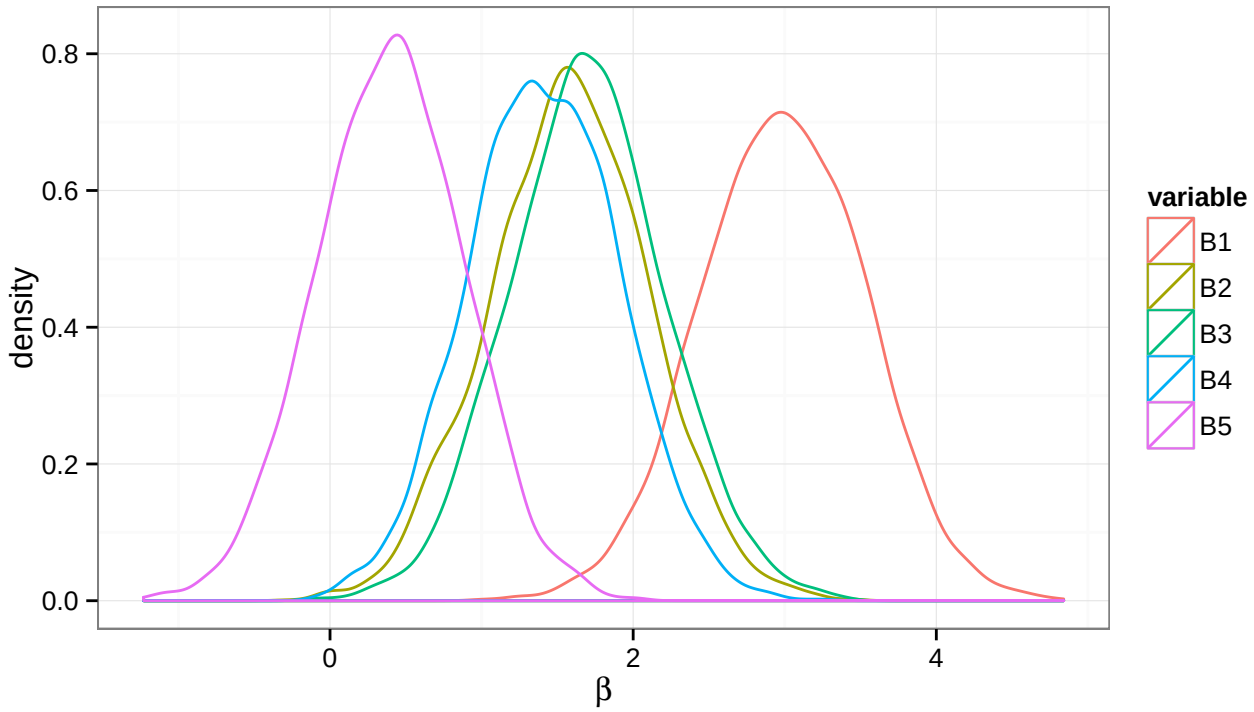

Figure 3: Densities for the posterior distributions of $\beta_1 \ldots \beta_5$ for the random intercept model. This plot has a similar shape as the fixed intercept model.

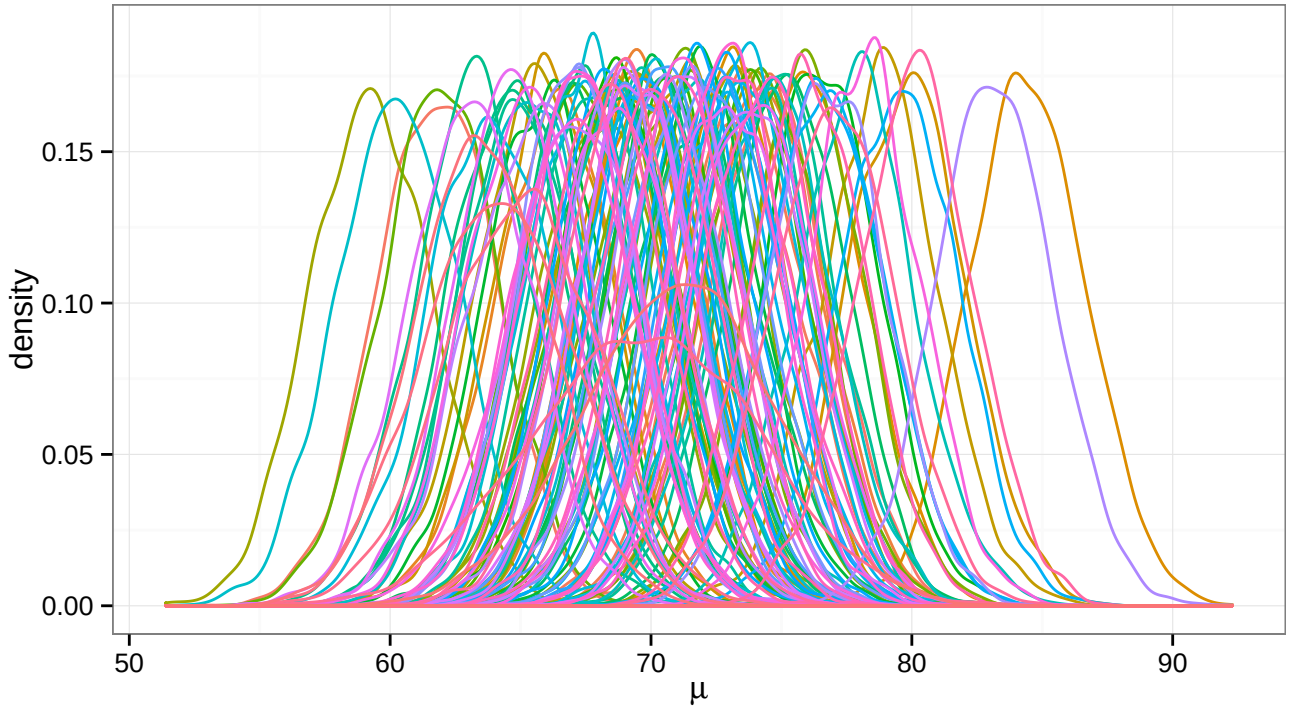

Figure 4: Densities for the posterior distributions of μ_i for $\{\mathrm{i} = 1, \ldots, 122\}$. This is not meant to be informative about specific teams, only to give a general overview of the distribution of the μ_is.

1	Boise.State	84.4411887005126
2	TCU	83.1901862250357
3	Wisconsin	80.0882734166512
4	Brigham.Young	79.8398217975428
5	Oregon	79.4859663911219
6	Cincinnati	78.7729403828867
7	Utah	78.1257998515623
8	Missouri	77.8413057943964
9	Navy	76.9861344319924
10	Stanford	76.7799010824779
11	Oklahoma.State	76.6968276023021
12	Oregon.State	76.6872837227628
13	Kansas.State	76.5616193548179
14	West.Virginia	76.0067463659355
15	Baylor	75.9403113895548
16	Georgia.Tech	75.9372616051767
17	Louisville	75.2637675275503
18	Michigan.State	74.842474893367
19	Nevada	74.6707220114264
20	Air.Force	74.6072057579171
21	Arkansas	74.3738402764138
22	Houston	74.3429465828683
23	Virginia.Tech	74.2627866117231
24	Arizona.State	74.2215815314545
25	UCF	74.0117444756685
26	Mississippi.State	73.9410746422882
27	East.Carolina	73.9301441465339
28	Northwestern	73.8731417310995
29	Iowa	73.7734957258427
30	Texas.Tech	73.7414953717759
31	Northern.Illinois	73.6190729590746
32	Arizona	73.5801136485171
33	Alabama	73.2950756638211
34	Wake.Forest	73.2098801346892
35	Boston.College	73.0062126073568
36	Ohio.State	72.91467153102
37	Tulsa	72.913091081173
38	Penn.State	72.863230841572
39	Connecticut	72.8487786355028
40	Vanderbilt	72.7727511536916
41	Rutgers	72.7719985547329
42	Oklahoma	72.4695768747156
43	South.Carolina	72.2162375299069
44	Fresno.State	72.1802945221322
45	Kentucky	71.9202572633037
46	LSU	71.8029120903811
47	Pittsburgh	71.6727826731039
48	Nebraska	71.3637992907067

49	South.Florida	71.3061590310206
50	Hawaii	71.2754557362564
51	Clemson	71.2234196482045
52	Texas.A.M	71.217945141784
53	Utah.State	71.1835380774053
54	Western.Kentucky	71.0523616770829
55	Washington	70.9215639660561
56	Syracuse	70.620808548161
57	mu	70.5727620552528
58	Iowa.State	70.5561727429466
59	Central.Michigan	70.5051329679142
60	Louisiana.Tech	70.4654382642888
61	San.Diego.State	70.458328013632
62	Marshall	70.3494283396014
63	Washington.State	70.1951120725453
64	North.Carolina.State	70.1361071097238
65	Georgia	69.9869902281284
66	Rice	69.9654206840531
67	Appalachian.State	69.7549173792088
68	Bowling.Green	69.6658252775858
69	Ohio	69.598359459957
70	Arkansas.State	69.5934654505691
71	Minnesota	69.5746143651497
72	Auburn	69.5579620429699
73	Duke	69.5269315039926
74	Toledo	69.4312127849379
75	Troy	69.3836444567825
76	San.Jose.State	69.3737543603263
77	Ball.State	69.3724393443107
78	Wyoming	68.9037839537383
79	Temple	68.8202232738949
80	UCLA	68.7372746885252
81	Colorado.State	68.7092670553952
82	Kansas	68.6996938464319
83	Mississippi	68.6725102252684
84	Indiana	68.6610845521884
85	California	68.4851761663893
86	Louisiana.Lafayette	68.4534003435596
87	Purdue	68.4379850814014
88	Florida	68.3550278871597
89	Western.Michigan	68.1916609537872
90	Florida.State	68.1093860281612
91	Middle.Tennessee	67.7288283828917
92	North.Carolina	67.3751799222775
93	Virginia	67.2993923614573
94	Notre.Dame	67.2797059111174
95	Southern.Methodist	67.2562633791517
96	Louisiana.Monroe	67.221733585877
97	Texas	67.2166675628145

98	Illinois	67.1886443869617
99	Maryland	67.0981881897859
100	Southern.Miss	67.0615713561811
101	UTEP	66.9984361149501
102	USC	66.8908841727585
103	Army	66.3459742477299
104	Buffalo	66.1725535984353
105	Kent.State	65.9198992662809
106	Colorado	65.5093390367157
107	Tennessee	65.3782097500374
108	New.Mexico	65.2663413973133
109	UNLV	65.219806178361
110	FIU	65.0911311755968
111	Michigan	65.0468674503411
112	Miami..FL.	64.8522872063077
113	UAB	64.6766309644693
114	Memphis	64.563963538547
115	Florida.Atlantic	64.1008802480949
116	North.Texas	63.8445191884288
117	Massachusetts	63.4871333132378
118	Miami..OH.	63.2336509238031
119	Tulane	63.0043920292004
120	Idaho	61.9089831306563
121	Akron	61.7896803480592
122	New.Mexico.State	60.2971853276509

Table 3: Table of individual team intercepts, unlabelled and sorted in descending order. All of the distributions are unimodal symmetric and have centers ranging from roughly 60 to 80.