Hierarchical Models For College Basketball Prediction

Matthew Edwards, 1000551333*

4/19/2021

Abstract

Using data on regular season and playoff performance of Division I Men's college basketball teams, I fit a series of Bayesian Models designed to try and account for difficulty in playing schedule by estimating conference-specific effects. The data, containing every playoff team and their final position in the playoff tournament, are used to estimate the probability that a given team will advance through each of 8 possible playoff stages. Exploratory data analysis reveals evidence that certain conferences consistently outperform others, and thus of the models I fit, those with fixed effects on conferences see the best results.

1 Introduction

One of the most exciting annual sport events is the playoff tournament for American Division I college basketball, known informally as 'March Madness.' But beyond the game itself, there is a massive following of statisticians and machine learning researchers who try and predict the outcomes of games. Perhaps the most common targets are win/loss (binary), or the point spread. Although there is money to be made if you can build a high performing model in this space, the interesting aspect is the debate among basketball enthusiasts over exactly what makes a team championship-worthy.

Most approaches tried to date involve hand engineering large feature sets based on extensive domain knowledge Gumm, Barrett, and Hu (2015). Other methods involve matrix decomposition to identify latent performance features Hao Ji and Li (2016). Some have even gone so far as to attempt to transfer learning from the features of other sports (see Ruiz and Perez-Cruz 2015). However virtually all existing approaches fail in one key aspect - they exhibit poor performance in outlier detection. More specifically, they fair to predict what are known as **upsets** - the term used to describe a scenario in which the losing team appeared much better on paper than the victor (hence the Madness).

One of the most universally known difficulties in predicting whether a team is good relative to their peers is determining the difficulty in their playing schedule. Historically, many

*Department of Statistics, University of Toronto, mr.edwards@utoronto.ca

teams with excellent records lose very early in the playoffs simply because, during their regular season, they were not exposed to the talent that permeates most teams who make it to the championship tournament. Given that most teams play most of their games within their conference (a cluster of schools usually close together geographically), the rise of one or two good teams often increases the play for the entire conference relative to other schools. The hierarchical nature of playoff performance is common knowledge, but seldom used in predictive models. In this project, I will model the playoff progress of March Madness teams in a hierarchical setting, using conference and year fixed-effects to account for these conference-level differences, which serve as a proxy for schedule difficulty.

2 Data

In this setting, the response variable will be an integer indicating how far the given team progresses in the championship tournament. Traditionally, there are 10 possible values this variable could take, corresponding to No Playoffs, First 4, Round of 64, Round of 32, Sweet Sixteen, Elite 8, Final 4, Runner-up, and Champion (in that order).

The dataset I use is from the popular Kaggle data science competition website, where they host a competition to predict similar metrics for the NCAA basketball tournament every year. In this case, I use data for every year in 2013-19, and 2021 (note that 2020 was cancelled due to covid). Each year, the exact same number of teams participate (347) in the Division I college basketball season, so for the span of 2013-2019, I have 2455 total observations. Note, however, that the playoffs are highly selected and consistently structured. Each year, only 68 teams (of the 347) make the playoffs. In terms of our prediction, this makes the target variable very unbalanced. Of the 2455 overall observations, only 476 (7 x 68) of these actually have a label that does not correspond to the lowest level (no playoffs = 0). For each observation, I have the following features:

1. Team
2. Conference
3. Number of Games Played, and Won
4. Offensive and Defensive Efficiency
5. Own and Opponent Field Goal Percentage
6. Own and Opponent Turnover Percentage
7. Own and Opponent Offensive Rebound Rate
8. Own and Opponent two-point and three-point percentage
9. Own and Opponent Free Throw rate
10. Tempo (Possessions per 40 minutes)
11. Wins above bubble (minimum wins to avoid tournament exclusion)
12. Seed
13. Latest Postseason round (the target)
14. Year

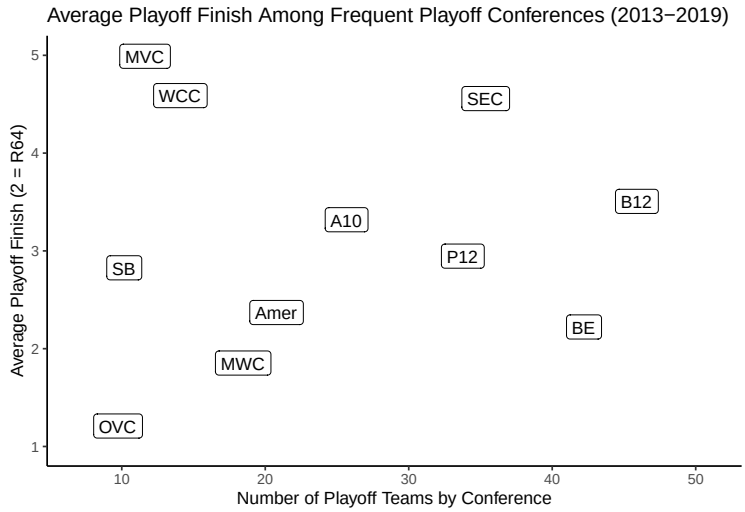
Including year, we have 22 features in total. Many of these features are fairly standard in models of prediction, however some (like wins above bubble, and defensive efficiency) are

advanced, and should provide valuable insight as to the veracity of many common beliefs (ie defense wins championships). One of the more interesting features is SEED, which is a numerical ranking assigned to each team, determined by expert committee near the end of the regular season. Many features like the ones I use here factor into the seeding process, however we know from experience that there is far from perfect correlation between a high ranking and playoff progress.

In terms of which variables act as controls, year and conference are the obvious choices. I may also try to control for SEED, however this is difficult to do well - it is known that even the expert committee often struggles with deciding the higher seeds (weaker teams), so at higher levels, this variable may lose predictive power. In addition to achieving good prediction, it will also be interesting to see which conferences rank highest in terms of fixed effects, and how this changes across time. Moreover, the 2021 tournament is currently underway, which gives statisticians a chance to test their models live. I should also note that this tournament has been around since the 1920s, but the current format is quite recent.

3 Analysis

The plot below shows the top 14 conferences in terms of the total number of team playoff appearances. In other words, it comprises only conferences that have more than one team make the playoff 68 in at least one year. The requirement of multiple teams in a given year is based on the fact that every conference has a within-conference tournament each year, and one spot in the playoffs is automatically awarded to the winner of each conference tournament. Over the years 2013-2019 (7 playoff tournaments), this guarantees that every conference has at least 7 total playoff teams. Thus the first plot shows conferences with 8 or more playoff entries.

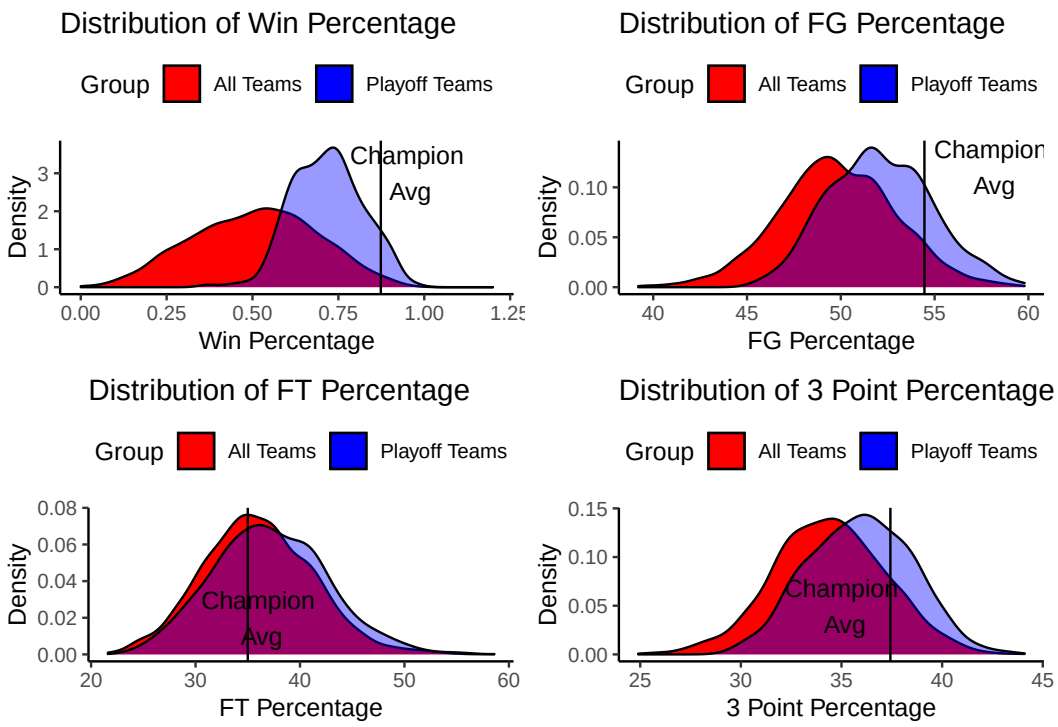


Looking at the number of playoff teams across this time frame, along with the average finish is quite revealing of the interconference differences in consistency. We can see that a small

number of conferences consistently send at least a handful of teams to the playoffs in most years (B12, B10, ACC), and many of these same conferences perform well above average when they arrive. Conversely there are several conferences that only rarely send more than one team, yet perform remarkably well compared to the rest of the field when they do appear. This presence of small handful of conferences that consistently outperform the field suggests that a model accounting for conference-specific effects could work well.

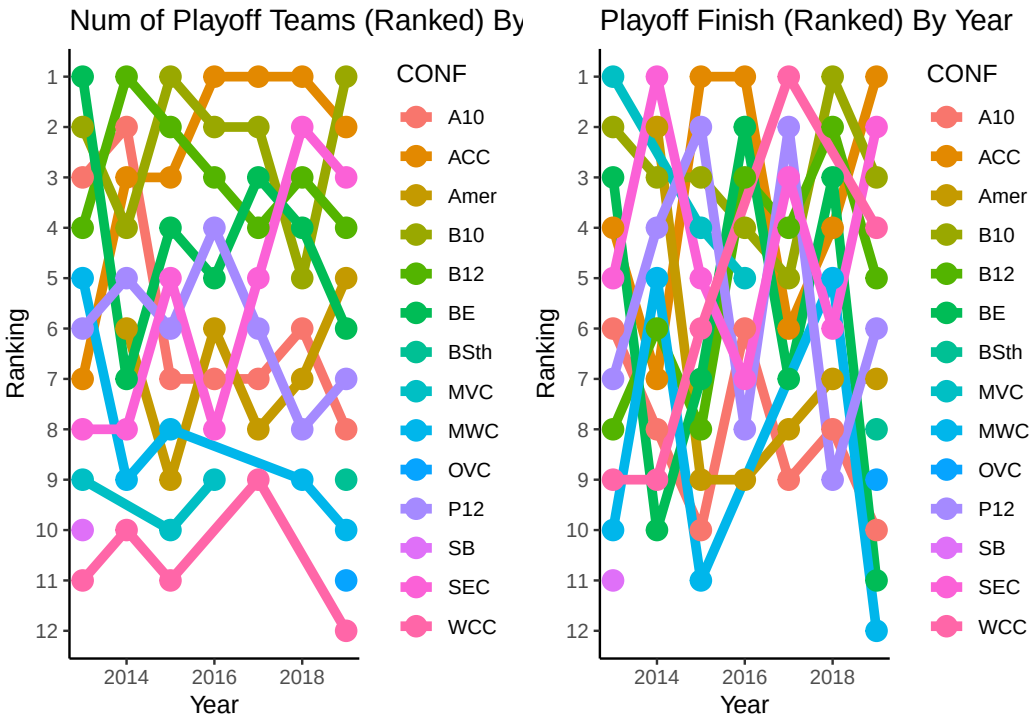
But conference is not the only predictor we have. Many of the post popular features in prediction tasks similar to this are traditional sports statistics. I described several such features in the data section, including wins, field goal percentage, free throw percentage, and turnover percentage. Intuititively we expect that Championship teams excel in these areas. However, given the bottom heavy structure of the playoffs, and the fact that many good teams do not make the playoffs (recall selection is decided by committee), we might wonder just how important these types of heuristics are in separating playoff teams from the others.

The following plot shows distributions of several key metrics for 1-all teams and 2- playoff teams, as well as the average for the championship team across all years.



This second figure confirms that these features, though not perfectly predictive of who wins the tournament, are definitely important. Though it is not necessarily true that the champion will be the best in any one, or even any one group of these metrics, we can see that, on average, the champions of the tournament are among the best in each of them, and that more generally, teams who make the playoffs tend to be above average in most of them. Though I have only shown these four (among the most popular) statistics, other works have shown that the remaining features also play important roles in determining team quality (see Fazelinia, Annamoradnejad, and Habibi 2020).

One last dimension that we might think plays a role in team outcome is time. Though we only have seven years worth of data (plus an eighth -2021 - for testing), it is worth examining how much of an impact year plays on conference and individual effects. Recall the earlier plot, containing the 14 conferences who sent multiple teams to the playoffs in at least one year. Overall there are 32 conferences in Division I, meaning 18 conferences have sent just the minimum number of teams in the years of our analysis. It is worth examining how these top conferences change by year.



We can see that, even among the best conferences, there is still variance year to year, although that same handful of conferences often finishes in the top 5 according to both number of playoff entries and average length of playoff run. This is not evidence of any overwhelming time-specific effects that may impact our models, but the above plots do show that (especially in this limited dataset) time is an important factor that should be included in our set of controls. Given that we clearly have both conference-specific effects and time-specific effects that may also depend on conference, a hierarchical model appears to be a promising approach. Let's not forget the importance of the heuristic features - it is likely that the best model will mix both hierarchical controls and a wide variety of basketball-related performance metrics.

4 Modeling and Results

Earlier analysis suggests that it is appropriate to include both conference-specific and year-specific effects in our model. We should also use many if not all of the covariates listed in

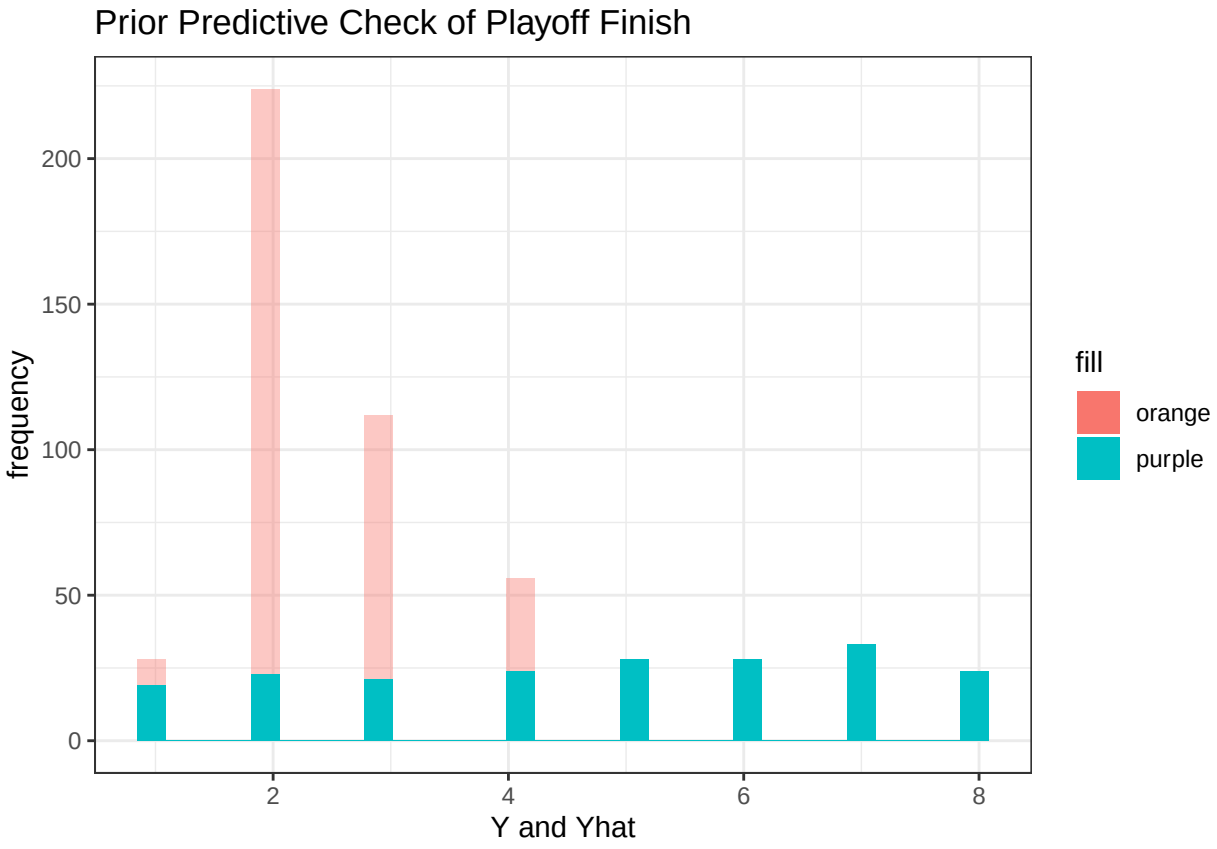
the data section. However it is not clear which of the two components is more important in terms of prediction. With this in mind, I fit 3 different models:

1. Softmax Regression with Covariates
2. Bayesian Hierarchical Model with Fixed Conference Effects
3. Bayesian Hierarchical Model with Fixed Conference and Year Effects

Note that each model must allow for the possibility that any team (among 347) makes the playoffs, even though for each year, only 68 actually do. Before fitting any models, I convert all index variables(team, year, and conference) to integers, and standardize all features.

4.1 Prior Predictive Checks

Note that in each of these models, I assume the standardized priors have coefficients with priors of $N(0, 1)$. To validate this assumption, the following plots show prior predictive checks for several key features. I run a thousand simulations of the coefficients on each feature. I then use each simulated value, along with the data to generate 1000 predicted values. Lastly, I compare the distribution of predicted values to the true labels.



We can see that under our prior parameters, we slightly overpredict the higher values. The natural shape of the posterior is exponentially decaying, which just reflects the pyramid structure of competition within the tournament. Unfortunately given high information in

the posterior, the priors would have to be very informative too, and in doing so we might end up with a very miscalibrated model.

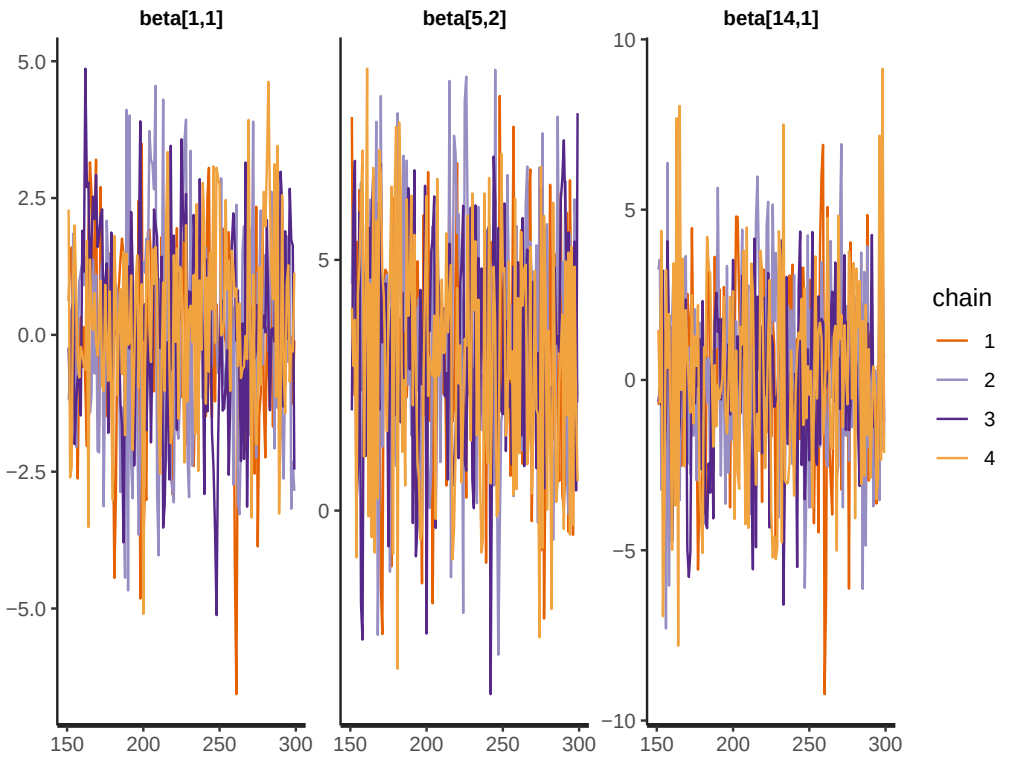
4.2 Softmax Regression with full covariates

This first model takes the following form:

$$\begin{aligned} \pi_i &= softmax(\beta_0 + \sum_{d=1}^{D} \beta_d X_{di} + \epsilon_i) \\ y_i|\pi_i &\sim Categorical(\pi) \\ \beta_d, \beta_0 &\sim N(0, 5) \\ softmax(t_j) &= \frac{exp(t_j)}{\sum_{k=1}^{K} exp(t_k)} \end{aligned}$$

Although this model could have been run in R, I run it using Stan for consistency across models. In this case, D = 17 (all features excluding year, conference, and the target), and recall K = 8. Note that the categorical distribution is just the multivariate version of the Bernoulli. The resulting coefficient matrix is $D \times K$, and the argmax of the columns gives the prediction. I use the default of 4 posterior chains, each running for 1000 iterations (half for warm up, and half for sampling). Please refer to the accompanying `softmax.stan` file for the code that runs this base model in Stan.

Though I won't go into great detail about the fit of this base model, the following traceplots show the mixing of the chains, and the eventual convergence. I plot the intercept, along with the coefficients on own rebounding and defender three-point percentage:



What we see is both good mixing and good convergence from each of the traceplots. This is an indication that the model is reasonably well-fit. But let's see how much better we can do by including fixed-effects for conferences.

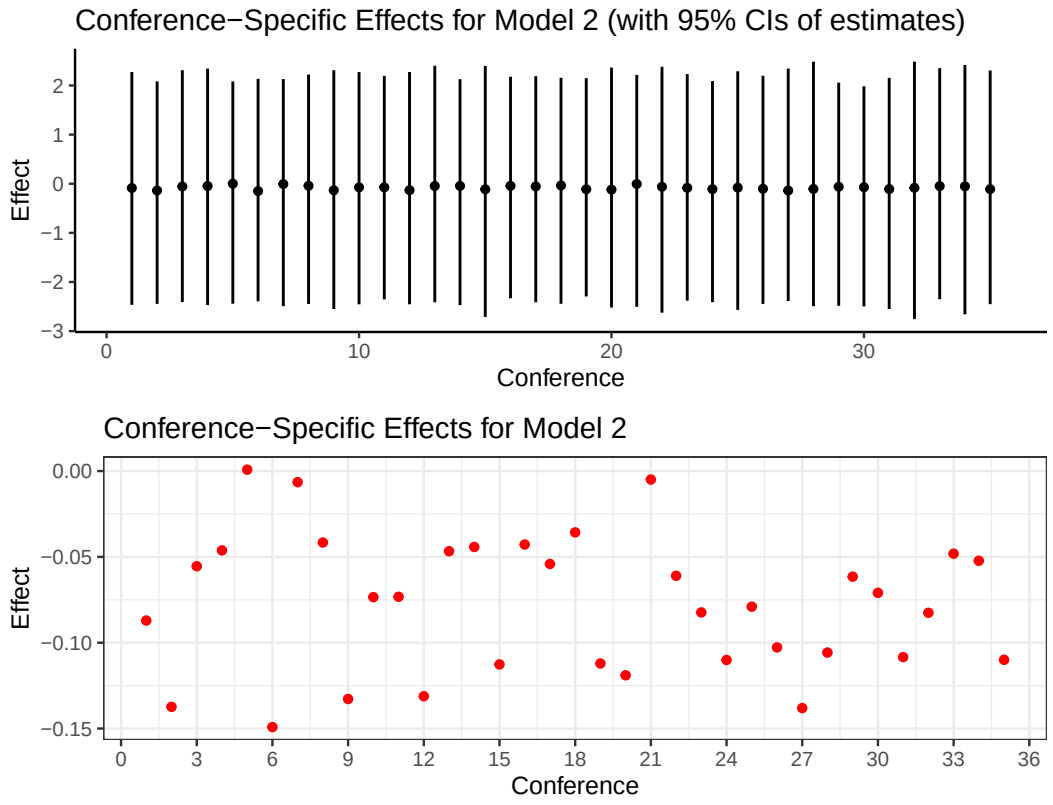
4.3 Bayesian Hierarchical Model with Fixed Conference Effects

This second model is similar to the baseline, only we add to it the conference-specific effect, which gets its own set of priors:

$$\begin{aligned}
\pi_i &= softmax(\beta_0 + \sum_{d=1}^{D} \beta_d X_{di} + \alpha_{j[i]}^{conf} + \epsilon_i) \\
y_i|\pi_i &\sim Categorical(\pi), i = 1, \cdots, N \\
\alpha_j^{conf} &\sim N(\mu_\alpha, \sigma_\alpha^2)\ j = 1, \cdots, J \\
\beta_d, \beta_0 &\sim N(0, 5) \\
softmax(t_l) &= \frac{exp(t_l)}{\sum_{k=1}^{K} exp(t_k)}
\end{aligned}$$

In this case (for time) I fit only 600 iterations instead of 1000, though I still use 4 chains. We still have D = 17, and K = 8, but now we also use J = 35 conferences. Please refer to the `softreg-conf.stan` file for the code to this model. Note that there are actually 32

conferences, but 3 teams in Division I are considered independent, and so are classified as individual conferences. Here the fixed effects are given normal priors (posterior predictive checks will confirm the feasibility of this is the next section), with the fixed-effect parameters having their own standard normal priors. The following plot shows the fixed effects by conference, along with standard 95% CIs:



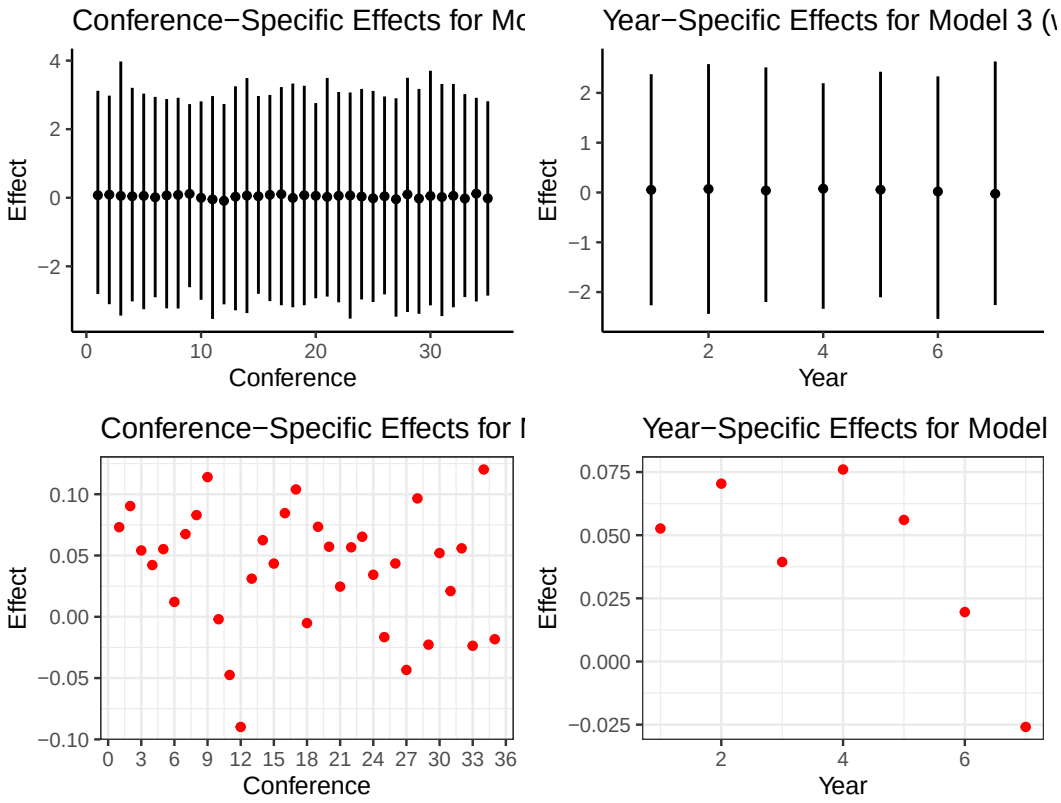
In the plots above, we can clearly see that the fixed effects are very noisy, and virtually all of them bracket 0, indicating they are not statistically significantly different from 0. However (p-values are not everything), looking at the values, we can see that a small number of conferences have the largest effects. In particular, conferences 5, 7, and 21 appear the highest. These numbers are linked to the Atlantic Sun, B12, and MEAC conferences respectively. These conferences contain teams that routinely go to the playoffs. This is not to say that most teams in those conferences earn spots every year, however recent playoff tournaments have contained a small (or in the case of B12 large) host of teams from them. Also of note is the high effect given to conference 18 (Ivy), which contains many schools that have only begun to recently make the playoffs regularly (Yale, Harvard, Cornell, etc.).

4.4 Bayesian Hierarchical Model with Fixed Conference and Year

This third model includes the second model, but now we add a year-specific effect also:

$$\begin{aligned}
\pi_i &= softmax(\beta_0 + \sum_{d=1}^{D} \beta_d X_{di} + \alpha_{j[i]}^{conf} + \gamma_{t[i]}^{Year} + \epsilon_i) \\
y_i|\pi_i &\sim Categorical(\pi), i = 1, \cdots, N \\
\alpha_j^{conf}, \gamma_t^{Year} &\sim N(\mu_\alpha, \sigma_\alpha^2)\ j = 1, \cdots, J\ t = 1, \cdots T \\
\beta_d, \beta_0 &\sim N(0, 5) \\
softmax(t_l) &= \frac{exp(t_l)}{\sum_{k=1}^{K} exp(t_k)}
\end{aligned}$$

As above, I use 600 iterations instead of 1000, but still 4 chains. In this case, the parameters are identical to those above, but we now have a second fixed-effect controlling for year (T = 7). The following plots show the year and conference fixed effects for Model 3. Please refer to the `softreg-conf-yr.stan` file for the code to Model 3.



These plots are very interesting, because once we control for the year of play, we see that many more conference-specific effects are positive, which is what we would expect given the fact that we know of at least a handful of elite conferences whose teams benefit from increased schedule difficult. Although the largest affects to not match directly with the

conferences I singled out in Model 2, we can see that some specific sequences of conference numbers (e.g. 3-6, mid-teens) correspond to the highest effects in both cases. This is not a coincidence - those numbers correspond to some of the oldest, most decorated basketball programs in the country, as well as some of the most established conferences. Unfortunately, we again see that the fixed effects themselves are quite noisy for both year and conference.

One last interesting thing to note is the recent decline in year effects, which on its surface might indicate a decline in play in recent years. This is difficult to assess, since playoff performance itself must also be measured relative to competition, so I won't speculate on whether those results are robust.

5 Conclusions

Although the models I fit above contain many controls that do not appear significant, they do match match reality in terms of conference effect assignment, and they do match the findings of my EDA. We can see (particularly based on the prior predictive check) that these models are likely to overpredict deeper runs into the playoffs. In this sense they are overly-optimistic. However, given the fact that schedule difficulty is moderately controlled, testing may reveal that certain types of predictions (here I'm thinking those in the middle rounds of the playoff tournament) are much improved. In order to know this, I would have to get new test data.

In a week or two the 2021 dataset (March Madness was cancelled in 2019 due to COVID-19) will be released, and would be great for testing purposes. Although there are again 347 teams involved in Division I, I would filter to only those teams who made the playoffs, since any loss I might be interested in requires the gold labels. Unfortunately, since the 2021 data was not available at the time of this writing I was unable to account for that year in the fitting of the models above. Thus test comparison would only valid for the first and second models we fitted (there is no estimate for the year-effect 2021). Comparing these models on recent data would give a more definitive answer in terms of quality.

References

Fazelinia, Amir, Issa Annamoradnejad, and Jafar Habibi. 2020. "Using Experts' Opinions in Machine Learning Tasks." http://arxiv.org/abs/2008.04216.

Forsyth, Jared, and Andrew Wilde. 2014. "A Machine Learning Approach to March Madness." Working Paper. Brigham Young University Department of Computer Science.

Gumm, Jordan, Andrew Barrett, and Gongzhu Hu. 2015. "A Machine Learning Strategy for Predicting March Madness Winners." *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 1–6.

Hao Ji, Adam Boudion, Erich O'Saben, and Yaohang Li. 2016. "March Madness Prediction: A Matrix Completion Approach." Working Paper. Old Dominion University Department of Computer Science.

Kvam, Paul, and Joel S. Sokol. 2006. "A Logistic Regression/Markov Chain Model for NCAA Basketball." *Naval Research Logistics (NRL)* 53 (8): 788–803. https://doi.org/https://doi.org/10.1002/nav.20170.

Ruiz, Francisco J. R., and Fernando Perez-Cruz. 2015. "A Generative Model for Predicting Outcomes in College Basketball." *Journal of Quantitative Analysis in Sports* 11 (1): 39–52. https://doi.org/doi:10.1515/jqas-2014-0055.