Applying Bayesian Hierarchical Regression to Analyze how Division and Sport Gender Ratio Impact Revenue in NCAA College Sports

Ben Aoki-Sherwood, Parker Johnson, Teagan Johnson

Abstract: In this study, we apply a Bayesian hierarchical multiple regression model to analyze whether and how division and sport gender ratio impact the revenue of NCAA college sports teams. If these results were significant, we calculate how significant each impact was. NCAA Division I, II, and III teams in 2019 that had both men and women participants were included in the study. We find that, all else being equal, Division I and II teams made significantly more revenue compared to Division III teams, while the gender ratio of a team was found to not have a significant effect on revenue. Also, we determine which schools tend to make more or less revenue compared to similar schools.

I. Introduction

College sports is a massive market in the United States, generating billions in revenue each year. In recent years, the revenue generated by sports teams has been at the center of a growing debate about whether college athletes should be paid. Van Rheenen et al. describes the basis of this controversy, saying there may be evidence of economic exploitation but that many people believe athletes are fairly compensated (Van Rheenen, 2013). Although we don't explicitly explore this discussion, our analysis provides some context for how much revenue colleges are making through sports. We are specifically interested in the difference in total revenue between different schools and different athletic teams. In particular, we analyzed whether a college's athletic division level or its gender ratio of athletes impact their sports teams' revenues, and if so, how significant of an impact each has. Mowad et al. claims that wage differences between men and women in sports are significantly unjust (Mowad, 2019). Our analysis tangentially explores this relationship, along with analyzing how gender ratio impacts a college's sports revenue. In this analysis, we implemented a Bayesian hierarchical multiple linear regression model to evaluate whether a college's division level or its gender ratio of athletes impact their sports teams' revenues, and if so, by how much.

II. Data

The data used for this analysis was from EADA (Equity in Athletics Data Analysis), which is a part of the US Department of Education. The data was collected from a survey by Equity in Athletics, but corrections made near the end of a year were not reflected in the data set. This is a mandatory survey that all institutions must submit to the Department of Education via their online survey (Benedictine College, 2022). We found this data from the TidyTuesday

project, which uploads various data sets weekly to their Github page; this data set was uploaded by Tom Mock (Mock, 2022). The relevant variables in the data set included the following:

- Year: The year the data is from
- Institution name: The name of the school
- Classification name: What division the school is in
- Participation men: The number of men on the specified team
- Participation women: The number of women on the specified team
- Total revenue for men and women: The total revenue for both men's and women's teams for that year
- Sport: The sport that team plays

The original data set contained 132,327 observations, but we subsetted the data to include only dual-gender teams from the year 2019 that were from either NCAA Division 1, 2, or 3 schools, which left us with 5843 observations. These observations came from 1,063 unique institutions and 31 different sports, with the most common sports including basketball, track and field, soccer, swimming & diving, and tennis. The number of participating men ranged from 1 to 192 men on a team, the number of women participating ranged from 1 to 211 women on a team, and the total revenue ranged from 2,297 to 42,177,260 US dollars. We also included a column for the ratio of men to women participants on each team, which ranged from a ratio of 0.016 to 17.0 men to women.

III. Methods

A. Model

We implemented an uninformative Bayesian hierarchical multiple regression model to assess the impact of sport team gender ratio as well as membership in NCAA Divisions 1, 2, or 3 on total sport team revenue for each school in our dataset. The model we created was defined by

$$Y_{i,j} \sim N(\mu_{i,j},\ 1/\sigma^2)$$

$$\mu_{i,j} = \beta_0 + \beta_1(D1_j) + \beta_2(D2_j) + \beta_3(x_{i,j}) + u_j$$

$$u_j \sim N(0,\ 1/\tau^2)$$

$$\beta_0, \beta_1, \beta_2, \beta_3 \sim N(0,\ 0.01)$$

$$1/\sigma^2, 1/\tau^2 \sim Unif(0,\ 100)$$

where $Y_{i,j}$ is the log revenue of sport *i* at school *j*, $\mu_{i,j}$ is the expected revenue of sport *i* at school *j*, σ is the residual standard deviation of sport revenues, and β_0 is the intercept parameter. $D1_j$ and $D2_j$ are indicator variables encoding whether school *j* is a member of NCAA Division 1 or 2, so the baseline when both indicators are 0 represents a school in Division 3. Thus β_1 is the expected difference in revenue between otherwise identical teams in Division 1 and Division 3, and β_2 is the same for Division 2 and Division 1. $x_{i,j}$ is the ratio of male to female athletes in sport *i* at school *j*, β_3 is the expected increase in sport revenue per one point increase in this ratio, u_j is a random effect of school *j* accounting for differences not captured by division or gender ratio, and τ is the standard deviation of this random effect. Note that all normal distributions in the model were parameterized in terms of mean and precision instead of mean and variance. Because of the massive range of possible sport revenues ($2,297 - $42,177,260), we used log

revenue as our response variable, allowing us to set uninformative priors with precision 0.01 for all of our regression coefficients.

Non-hierarchical linear regression assumes that the data has a constant variability about the regression line and that observations are i.i.d. normally distributed at each set of explanatory variable values. Because schools in any given division and with any given sport team gender ratio may have access to vastly different fan bases and sports markets and thus may generate different revenues, we introduced a school random effect to avoid having to make the second assumption listed above. Because our model was a regression model, it also relied on the assumption of linearity between each explanatory variable and the log sport revenue, our response. Exploring the data, we found that there appeared to be a constant offset in the log revenue distributions of schools in each division, indicating that this assumption was a reasonable one to make. Finally, we exponentiated each parameter of our model to interpret these coefficients as multiplicative factors in sport revenue instead of linear slopes for log sport revenue.

B. Computation

We used JAGS to run 3 chains of MCMC to fit our model, with 1000 iterations of adaptation, 5000 iterations of burn-in, and 5000 iterations of sampling. To assess convergence, we first examined the trace and ACF plots for our sampler. In our first attempt at fitting with no thinning, the ACF did not decay quickly enough for several parameters, so we re-ran the MCMC sampler with a thinning interval of 10, leading to a suitably efficient sampler. To verify the sampler's efficiency, we checked that the effective sample size for all parameters was at least 2000. To verify that the simulation converged on a stable posterior distribution, we ensured that the Gelman-Rubin statistic for every parameter was at most 1.1, and also calculated the Geweke

statistic for every parameter, checking that none of these statistics was too far from 0. After confirming that our model converged, we performed posterior predictive checks to ensure that the data could plausibly have been generated from the data.

IV. Results

Table 1 below displays 95% probability intervals for the exponentiated regression coefficients from our model. These probability intervals indicate that there is a 95% probability that a Division III sports team with all women and no men and no random effect multiplier will generate between $127,902.7 and $150,446.9 in revenue, there is a 95% probability that, all else being equal, a Division I sports team will make between 7.36 and 8.77 times more revenue than the equivalent Division III team, and a Division II team will make between 3.05 and 3.64 times more revenue than the equivalent Division III sports team, and there is a 95% probability that an increase of one in the male-to-female gender ratio will correspond to a multiplicative increase of between 0.94 and 1.03 in sport revenue.

Parameter / Posterior Percentile	β_0	β_1	β_2	β_3
2.5%	127902.7	7.361201	3.048971	0.9421152
97.5%	150446.9	8.771056	3.643962	1.0332656

Table 1: lower and upper bounds of 95% probability intervals for each of the exponentiated regression coefficients in our model.

The distribution of school random effects is shown in Fig 1 below. The actual values of these effects can only be interpreted relative to each other, so we have labeled the schools with the top 3 and bottom 3 random effect values. We used the median of the posterior distribution of

random effect values as a point estimate for each school's random effect.

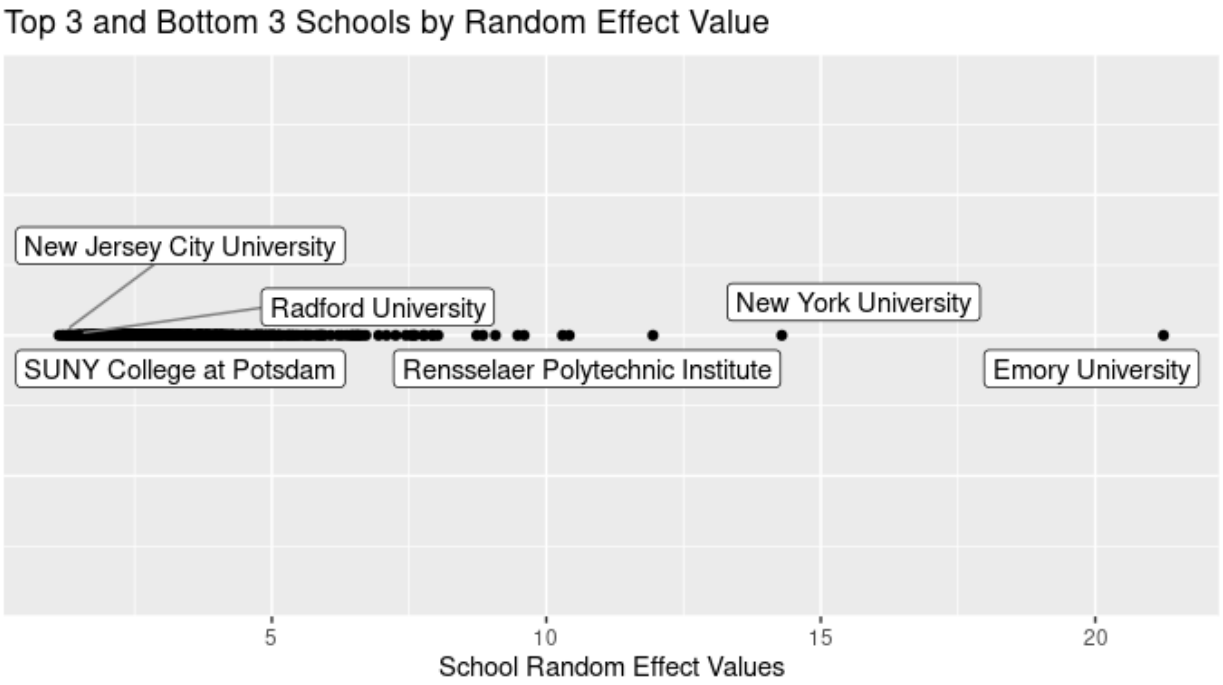

Figure 1: The distribution of random effect (u_j) values by school, with the top 3 and bottom 3 schools labeled.

As can be seen in Figure 1 above, the top 3 schools that generated unusually high revenue within their given division and gender ratio were Emory University, New York University, and Rensselaer Polytechnic Institute. The bottom 3 schools for this metric were New Jersey City University, SUNY College at Potsdam, and Radford University. The random effect values for these schools are shown below in Table 2.

School	Emory University	New York University	Rensselaer Polytechnic Institute	Radford University	SUNY College at Potsdam	New Jersey City University
Random Effect	21.24	14.29	11.94	1.21	1.18	1.12

Table 2: The three schools with the largest and smallest random effects in our model.

V. Discussion

We are 95% certain that the multiplicative increase in revenue in response to an increase in the gender ratio captures unity, so we conclude that the gender ratio of a sports team does not have a statistically significant effect on the total revenue of that team. Since the 95% probability interval for the multiplicative increase in revenue based on division was significantly greater than 1 for both Division I and II, this means Division I and II schools generate much more revenue on average compared to Division III teams, which could allow Division I and II teams to develop better programs with a larger budget. This makes sense given that schools in the top two divisions are allowed to recruit athletes on scholarship, leading to generally higher-quality athletes, and such schools also tend to have larger student bodies and influences than schools in the smallest division. These results could provide support for monetary compensation of Division I (and possibly Division II) athletes given the large amounts of money these athletes generate for their teams. As seen in Table 2, the schools with the largest random effect values are Emory University, New York University and Rensselaer Polytechnic Institute. Future work could be conducted to determine why these schools generate such high revenues as compared with schools in their respective divisions. Similarly, future work could be conducted to determine why Radford University, SUNY College at Potsdam, and New Jersey City University generate significantly lower revenues than schools in their respective divisions.

It is apparent that sports programs across all divisions generate significant revenue for their respective institutions. While this data set did not contain any information about the success of programs, it's likely that the more successful programs will generate more revenue. Future research would need to be conducted to assess this possibility, but if this were the case, then further advocacy could be given for compensating successful programs. Regardless, we have

shown how much more money Division I and II programs make compared to Division III, and which schools are generating unusually high and low amounts of revenue. While a statistically significant difference based on gender ratio was not found, further research would need to be conducted explicitly comparing the revenue of men's versus women's programs.

One drawback to this design is the possibility for inaccuracies to arise when justifying the prior distribution and model. While a weakly informative prior does not have a large impact on the posterior distribution, it does sway it slightly, so an incorrect prior distribution will result in a slightly inaccurate posterior distribution and interpretations. Since the value for β_0 ended up being much larger than the prior distribution that had a mean of 0, perhaps more research could have been conducted to find a more accurate prior distribution for each variable, especially the intercept parameter.

Another drawback to this design is that it does not categorize the data based on sport. Certain sports are likely to generate more revenue than others: sports such as basketball generate much more revenue on average than sports such as golf, yet this model does not capture this as we instead focused on categorizing the data based on division and school instead. This could explain why the random effects of the schools were so varied, as teams with large fan bases in very profitable sports like basketball would generate much more revenue for their schools. Future research could be conducted based on the sport of the team to determine which sports earn more revenue than others. With this, another drawback to our design is that teams that only consisted of one gender were not included in our analysis, so inferences about sports such as football and (in many cases) volleyball could not be drawn. While it's likely that men's teams generate more revenue on average than women's teams, our model does not take into account

revenue differences between mens' and womens' teams, so a future study incorporating these differences might help to address these issues.

VI. References

1. *Equity in Athletics (EADA) report*. Benedictine College. (n.d.). Retrieved November 18, 2022, from https://www.benedictine.edu/about/gov/eada/index
2. Mock, T. (2022, March 29). *rfordatascience/tidytuesday*. GitHub. Retrieved November 19, 2022, from https://github.com/rfordatascience/tidytuesday/tree/master/data/2022/2022-03-29
3. Van Rheenen, D. (2013). Exploitation in college sports: Race, revenue, and educational reward. International Review for the Sociology of Sport, 48(5), 550–571. https://doi.org/10.1177/1012690212450218
4. Jad Mowad (2019): Gender Inequality in Sports, Fair Play. Revista de Filosofía, Ética y Derecho del Deporte, vol. 13, p. 28-53