The Effect of Academic Performance on Athletic Success in Collegiate Athletic Programs

Derek Brickley

Faculty Advisor: Professor Andrew J. Sage

Lawrence University

3/16/2022

Introduction

Across the country, there are more than 500,000 college athletes competing for 1,100 different schools in the National Collegiate Athletics Association (NCAA). These athletes attend their institution not just to play their respective sport and succeed athletically, but also to pursue a degree and achieve their academic goals. With college sports being over a $18 billion business, why would the NCAA and college programs emphasize the importance of academics when the NCAA and athletics do not benefit. After all, coaches scout high school athletes nationwide looking for the best talent to bring into their individual programs. However, in addition to improving the athletes' life after their athletic career, academics may also improve success within athletic programs. By using a mixed effect model, we can look at how success in the classroom is related to how a team performs. In my analysis I will look at Division 1 programs' Academic Progress Rate (APR), measuring the eligibility and retention of athletes, and win percentage from nine different sports from 2010 to 2018.

Intuitively, I expect to find that programs with higher academic performance will result in higher win percentages. A program with higher academic performance will be more attractive for prospective high school athletes in their search to further their athletic and academic careers. More competition at a school due to more attraction from increased academics will increase the skill levels of its athletes (Lawrence et al.). Furthermore, outstanding work in the classroom will directly show in outstanding work in the students' respective sports. Higher academic performing programs will have athletes that are better at studying and preparing. If a program's athletes consistently are able to prepare for tests and assignments, then they will give more effort into their game preparation. Students that have a desire and passion for continuous learning will undoubtedly do better in school than students that lack the same motivation. When these high performing students take this mentality to sports, they will be more coachable players and do better with the constructive criticism that comes with it (Battle). Such a player will strive to improve daily and use the teaching from coaches to do so. Also, these high performing student-athletes will want to learn more about every aspect of their sport, raising their awareness and improving performance.

Data

The data in this research merged two datasets from the NCAA and data from the College Scorecard. The resulting dataframe is panel data that contains data from 394 schools and nine different sports over eight years. This study focuses only on Division 1 team sports and the relationship between team academic performance and that team's athletic success. Specifically this data was filtered to focus on nine team sports: Men's Soccer, Women's Soccer, Baseball, Softball, Men's Basketball, Women's Basketball, Football, Men's Ice Hockey, and Women's Ice Hockey.

In total there are 17,700 observations with 17 variables. Table 1 (below) shows the first five rows of the data and the 12 main variables that are important to this research, excluding ID numbers, shifted variables for year and APR score and the specific state of the school.

Table 1: Selected variables from the first five observations in the dataset.

	School	Region	Percent Communication Major	Total Enrollment	Percent Male	Percent White	Average Cost of Attendance	Average Age at Enrollment	Sport	Win Percentage	Multi-Year APR	Year
1	University of Virginia	5	0	14,232	44.330	59.840	21,142	20	Baseball	82.400	986	2,010
2	Vanderbilt University	5	2.910	6,836	48.960	63.910	54,718	19	Baseball	81.800	996	2,010
3	University of South Carolina, Columbia	5	7.290	21,033	46.530	78.230	21,490	21	Baseball	79.700	953	2,010
4	Stony Brook University	2	0.880	16,044	52.500	38.230	18,319	21	Baseball	77.800	959	2,010
5	University of North Carolina, Chapel Hill	5	14.860	17,943	40.980	66.140	17,777	20	Baseball	76.100	969	2,010

The first dataset used is the NCAA Academic Progress Rate Database. From this the variables of interest are the schools, the state each school is in, the various sports, the individual years, and each specific program's APR score. Each athlete on the team can receive at most 2 points, one for remaining eligible over the semester and another for staying at the institution. So, each athlete can have anywhere between zero and two points for a given semester. The APR score for each Division I team is calculated by adding all points earned by the team and dividing that total by the maximum possible number of points that could have been earned. This percentage is then multiplied by 1,000 to return the team's APR score. Thus, an APR of 950 means that the student-athletes in the cohort earned 95 percent of the eligibility and retention points that they could have earned. It is important to note that APR in this dataset represents a four-year rolling average of APR scores. Figure 1 (below) shows the variation in APR scores between sports and years. Men's Basketball, Baseball, and Football all have lower APR scores on average from the other sports (a), where Men's and Women's Ice Hockey has higher APR scores on average. We can also see that over the 8-year time period in our data (b) APR scores on average have increased.

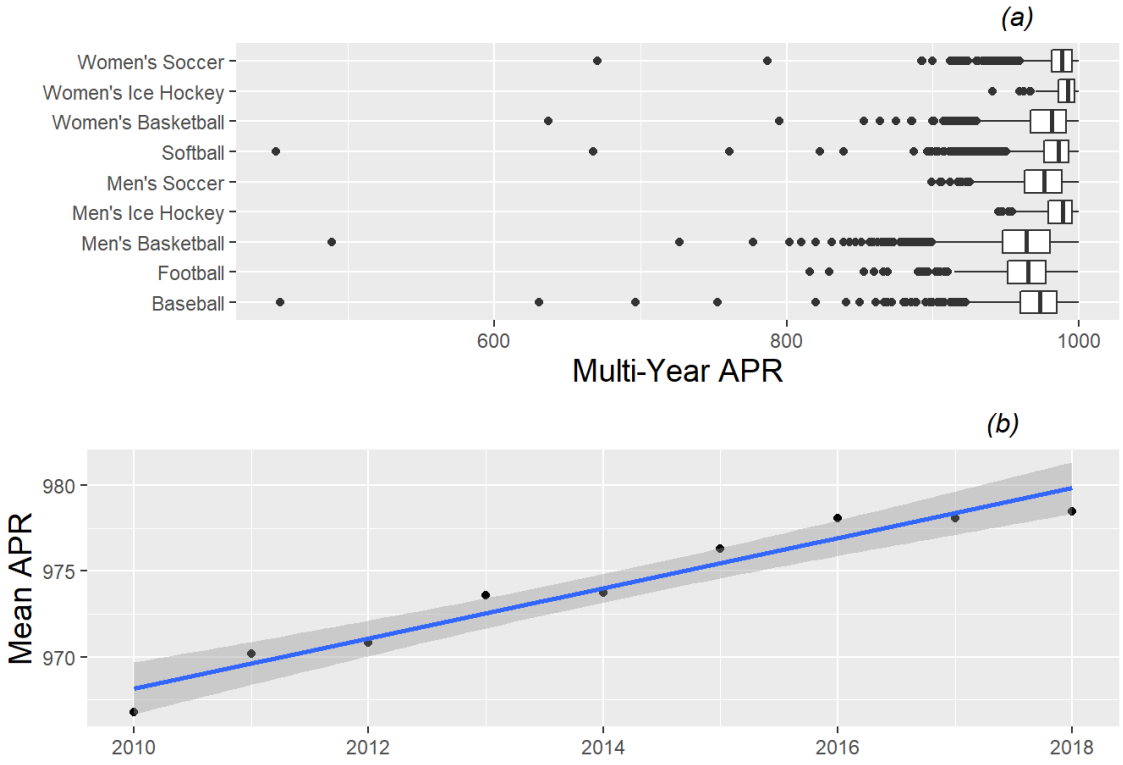

Figure 1: Distribution and trend of Multi-Year APR scores. Plot (a) is based on all 17,700 observations from all 2,168 programs over all years separated by sport. Plot (b) tracks the increasing trend of APR scores from the year 2010 to 2018.

The second NCAA dataset is a combination of datasets of the nine different team sports and their athletic performance in each year from 2010-2018. The important variables from this database are the team name, the team's win percentage, and the year of observation. With each dataset, it was necessary to create a new variable for the sport to keep track of which program is being observed when all datasets are merged together. In some cases, the individual sports reported win percentage differently than others. Multiple sports reported win percentage as a percentage with range from 0 to 100, while other sports reported it as a decimal 0.00 to 1.00. For uniformity and ease of interpretation, all percentages were changed to a scale of 0 to 100.

The final data accessed is the College Scorecard data. This database provides information about each specific institution not specific to athletics. Each observation has the school's ID number and the name of the school. For each school in each year the variables that were considered for their relevance are the geographic region of the school, the percentage of degrees awarded in Communication, Journalism, And Related Programs, the total enrollment, the percent of the student body that is white, the percent of the student body that is male, the average cost of attendance in each year, and the age at which students enroll at the institution. As with any form of modeling it is important to hold constant variables that may bias results and affect their validity. The College Scorecard data is used to include school-level variables and to determine what aspects of a school would have an effect on its sports' win percentages.

Table 2: Summary Statistics

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Percent Communication Major	5.4	4.0	0.0	2.8	4.9	7.7	23.3
Total Enrollment	13,404.3	10,200.8	947.0	5,099.2	10,173.0	19,606.0	59,183.0
Percent Male	46.0	7.9	0.0	41.6	45.6	49.8	93.2
Percent White	58.8	21.8	0.0	48.0	64.6	75.2	92.8
Average Cost of Attendance	32,295.2	15,949.0	10,605.0	19,988.0	24,441.0	46,143.0	72,960.0

Average Age at Enrollment	21.2	1.7	19.0	20.0	21.0	22.0	32.0
Win Percentage	50.1	17.7	0.0	38.0	50.0	62.5	100.0
Multi-Year APR	974.4	22.8	451.0	963.0	979.0	990.0	1,000.0

The first variable that was considered for its relevance was geographic region. In Minyong Song and Yang Zhang's Study, *Research on the Relationship between Geographical Factors, Sports and Culture*, they conclude that "The geographical environment influences the emergence and development of sports events in different ways, Climate changes will affect the conditions of sports, but also interfere with the athletes' body mechanism and emotion, thereby affecting athletic ability." Their findings give merit to consider geographic region as a possible influence on win percentage. The breakdown that follows is the classification of the geographic regions:

- 0 U.S. Service Schools
- 1 New England (CT, ME, MA, NH, RI, VT)
- 2 Mid East (DE, DC, MD, NJ, NY, PA)
- 3 Great Lakes (IL, IN, MI, OH, WI)
- 4 Plains (IA, KS, MN, MO, NE, ND, SD)
- 5 Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV)
- 6 Southwest (AZ, NM, OK, TX)
- 7 Rocky Mountains (CO, ID, MT, UT, WY)
- 8 Far West (AK, CA, HI, NV, OR, WA)

One other variable that needs to be considered is the percentage of the student body majoring in communication. This is a very specific variable, but its potential inclusion in our data stems from the phenomenon of major clustering. Programs and schools that cluster their student athletes into an easier, or less time-intensive major may have more energy to focus on athletics. In addition, this may allow more student-athletes to remain eligible with the smaller work-load that some majors may offer. Different institutions have been accused of clustering into different majors, but Schneider, et al. reinforce communications as one of the common majors subject to clustering. As such we would expect that as the percentage of the student body majoring in communication increases we would see a corresponding increase in those programs' win percentages. Kaydee McCormick found in their research that although clustering appears to have the possibility to occur among all college students, it seems to be more prevalent within intercollegiate athletics. As such, if we find clustering in the general student-body it is reasonable to assume that the athletic programs at that institution are subject to clustering as well. Although major clustering may be an issue in the data, there is no precise measurement of it, but this variable is our attempt to control for it. For uniformity and ease of interpretation, all percentages were changed from a scale of 0.00 to 1.00 to a scale of 0 to 100.

Three other variables of the student body are potentially relevant for this research: The percent of the student body that is white, the percent of the student body that is male, and the average age of enrollment. Analysis performed by Zippia and BBC show much variation within these specific demographics. Collegiate student-athletes are a sample of the general student body, and as such we assume the demographics of the student body are representative of their student-athletes. Ethnicity, gender, and age all impact sport participation, so we must consider them as possible controls. Again for uniformity and ease of interpretation, the percentage male and percent white were changed from a scale of 0.00 to 1.00 to a scale of 0 to 100.

The final institution level variable considered is the average cost of attendance. Studies by Battle, Ayers et al., and Gurney analyze the relationship between academics and athletics considering the extreme burden and time commitments that student-athletes have. Regardless of their concluding argument, they rely on this connection. With such extenuating circumstances and limited time commitments, athletes do not have time and are often not allowed to work full-time and only in 2007 did the NCAA enact a policy allowing student athletes to work part-time jobs during the off-season. With that said, other than financial aid many student-athletes cannot be expected to

afford the large tuition fee most institutions charge, not including student loans. Controlling for financial aid in this way adds value to the model by incorporating the affordability of different institutions, particularly for student-athletes who cannot, due to time restrictions, handle full-time employment.

In the merged dataset there were multiple of the included variables that needed to be modified. The first of these was the variable for each Academic Year. This variable was a strain that contained the two years (the start of the year and the end of the year) connected by a hyphen. In order to model over time, it was necessary to convert the year variable to an integer. The year used in this research is year from the start of the academic year. For example, the year 2010 is the 2010-2011 academic year. It was also important to rescale APR and years for later interpretation. From the APR score a new variable was created to demonstrate the centered APR score, or each observation's distance from the mean APR score of 974.4. The year variable also needed to be centered; a new variable is created where the first year in the data, 2010, is represented as 0 and years are measured as years since 2010. For example, the year 2018 would be represented as 8. There was also a need to create a unique identifier variable for each program, i.e. a specific sport team at a specific institution.

Methodology

Our main purpose is to interpret how program's APR scores, measuring academic success, impact program's win percentages. Using Ordinary Least Squares regression and multiple mixed-effects models, we can evaluate this relationship. However, it is important to include variables in our models that impact a program's win percentage outside of APR scores. By looking at each variable from the College Scorecard Database and its correlation with win percentage we can see if it is valuable to the models. Figure 2 (below) explores this correlation between each variable, where a darker red shading of the box and a more negative number indicates a stronger, negative correlation and a larger positive number with a darker blue shading indicates a stronger positive correlation. Looking at the column for win percentage in Figure 2, we can see that APR Scores, the percent white, and the total student enrollment are the most correlated with win percentage, the percent communication and percent male have very small correlation, and the average cost and average age have essentially no correlation. With no correlation, the average cost and average age of enrollment will not be included in the models as they add no value to our results. We will include models that include these variables, and models that do not to determine the effect these controls may have based on the differences in our results.

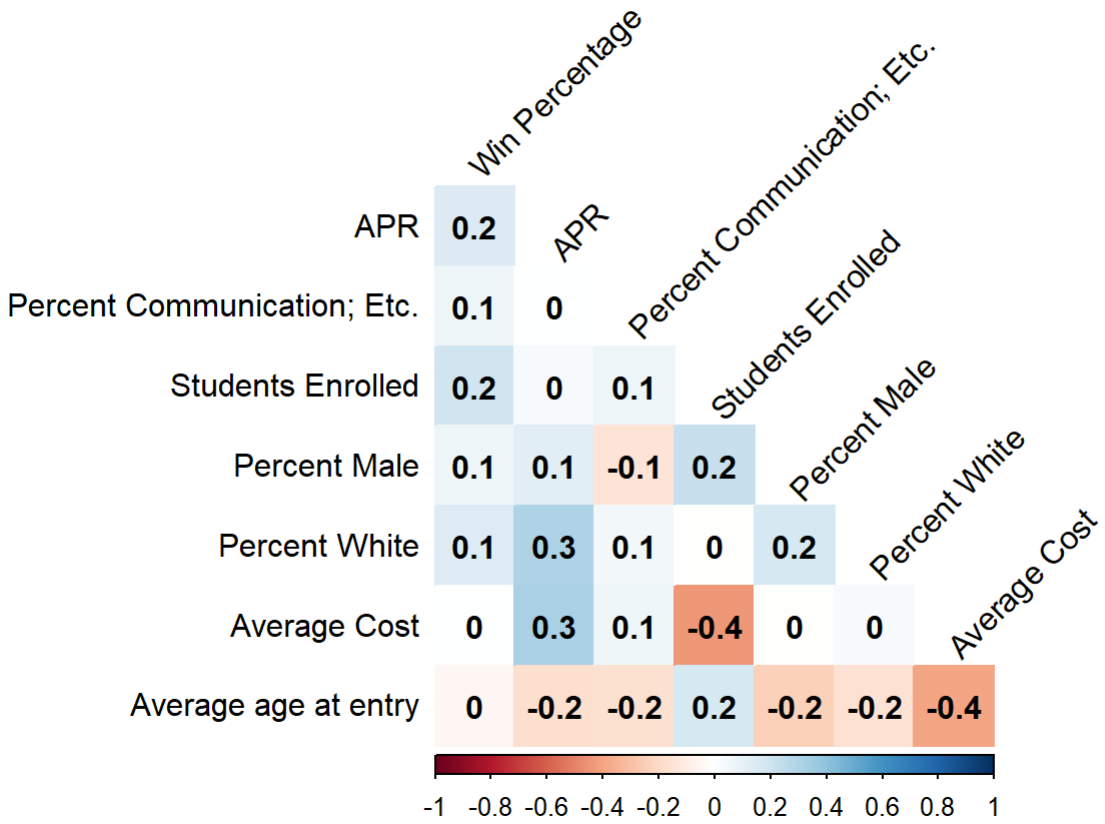

Figure 2: The correlation between variables

It is also important to explore the possibility of differences in location. Perhaps programs in the Southwest have higher win percentages on average than programs in New England. Figure 3 (below) demonstrates differences in win percentage for the different geographic regions. Generally speaking, all regions have a median win percentage of roughly 50%. Without major differences between regions, they may not contribute to our final results, but we will look at our models with and without a variable controlling for the geographic region and compare the estimates.

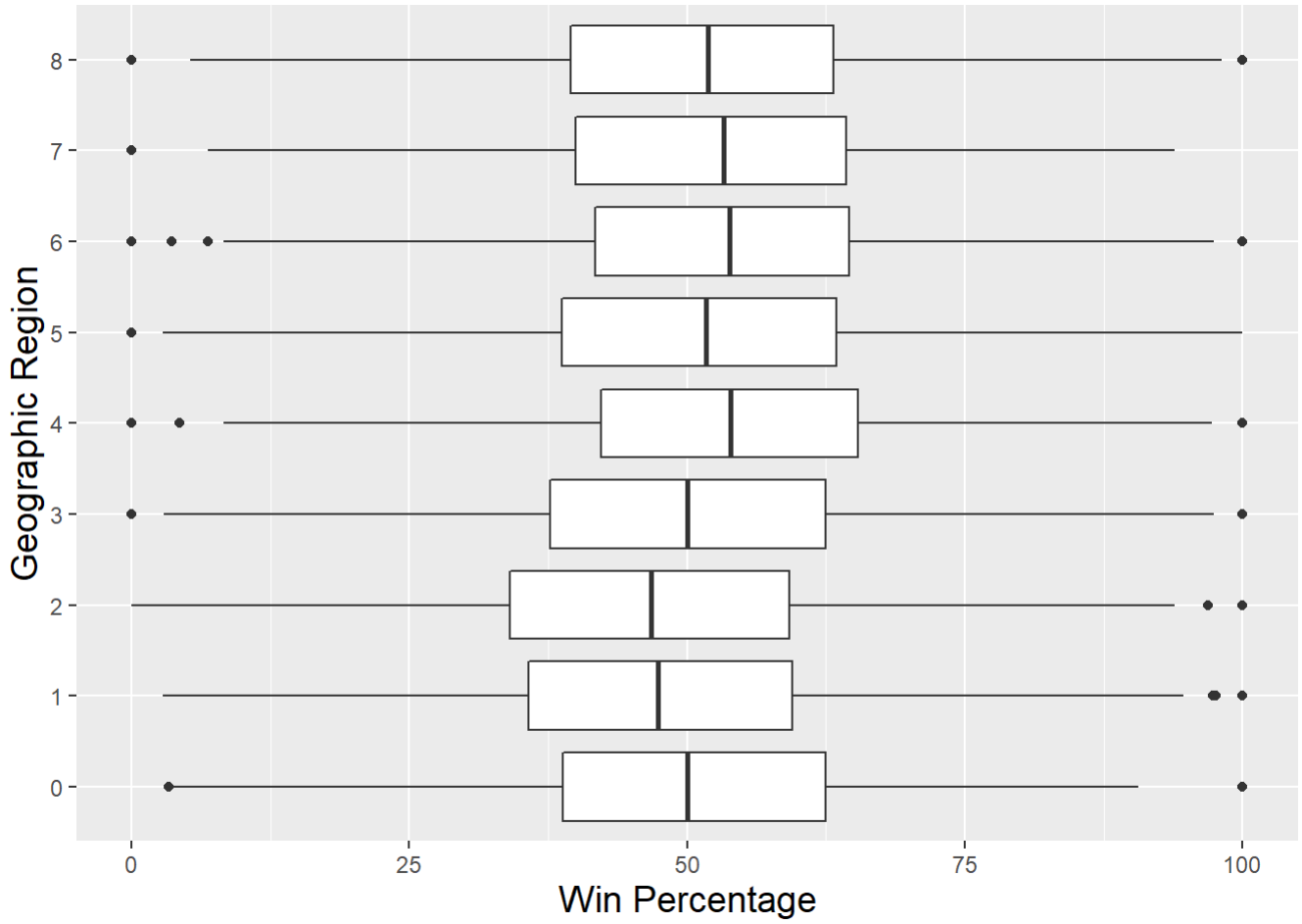

Figure 3: The distribution of Win Percentages in different geographic locations

To begin with we can look at an Ordinary Least Squares (OLS) regression of Multi-Year APR scores on Win Percentages. Model 1 below is the equation used in our research to calculate the OLS regression coefficients. Y_{ijk} represents the Win Percentage of program j at school i in year k. The β terms in this model are the regression coefficients corresponding to each variable in the model. ϵ_{ijk} then represents the error term capturing the unexplained variability in the model; the random difference of everything from its expected value. This error term is expected to be normally distributed where $\epsilon_{ijk} \sim N(0, \sigma^2)$ or in other words, everything deviates from its predicted win percentage randomly and independently. Model 1 is the standard OLS model using only APR scores to predict program's win percentage. The second model uses APR in addition to the most correlated variables from the College Scorecard Database. Model 3 depicts a model that also includes the variables that are very slightly correlated in addition to dummy variables for the different geographic regions. For OLS regression results, see Appendix Table 6.

Model 1: $Y_{ijk} = \alpha_0 + \beta_1 APR_{ijk} + \epsilon_{ijk}$

Model 2: $Y_{ijk} = \alpha_0 + \beta_1 APR_{ijk} + \beta_2 Enrollment_i + \beta_3 PercentWhite_i + \epsilon_{ijk}$

Model 3:

$$\begin{aligned} Y_{ijk} =& \alpha_0 + \beta_1 APR_{ijk} + \beta_2 Enrollment_i + \beta_3 PercentWhite_i \\ &+ \beta_4 PercentMale_i + \beta_5 PercentCommunication_i + \beta_6 NewEngland_i \\ &+ \beta_7 MidEast_i + \beta_8 GreatLakes_i + \beta_9 Plains_i + \beta_{10} SouthEast_i \\ &+ \beta_{11} SouthWest_i + \beta_{12} RockyMountains_i + \beta_{13} FarWest_i \\ &+ \beta_{14} Outlying_i + \epsilon_{ijk} \end{aligned}$$

However, our data violates a vital assumption of OLS models: that all entries in the data are independent. There are three levels in our model in which the entries have the potential to be correlated to one another: the school or institution level, the level containing each athletic program, and the year level. The key difference between an OLS model and a mixed effects model is the inclusion of random effects, variables that control for the random variability within each level in our model (Legler). The first level, the schools, are very important to consider. Any entries from a specific college or university have the potential to be correlated with other entries from the same school. For example, the win percentage of any program at the University of Michigan is going to be correlated with the win percentage at the University of Michigan in other entries, therefore not all entries are independent. The second level in our model is the program level. In this level it is important to take note that the win percentage of a specific program in one year is going to be correlated with other entries from the same program. The University of Alabama historically has a very successful football program, and we can expect that the win percentage of the University of Alabama in 2010 is not completely independent from its win percentage in 2011. The third and final level in this model is the year. For this level we need to consider that for any one program at any specific school, there will be random difference from year to year.

When using a linear mixed effects model, there are both fixed effects and random effects. In our models, the β terms represent the fixed effects. These variables will be constant over the program or school it represents. In our mixed effect models the random effects, that vary over each program or school, will be represented by s_i, p_{ij}, m_i, m_{ij} and $\epsilon_{i,j,k}$.

Instead of using least squares to estimate our regression coefficients, the linear mixed effects model uses restricted maximum likelihood, or REML. This separates the part of the data used to calculate the variance estimates from that used to calculate the fixed effects, producing an unbiased variance estimate (Oskolkov).

The first random effect our mixed effects model will include is s_i. This is a random effect allowing for deviation from each specific school after accounting for APR score and other fixed effects. Compared to the OLS model which only has the single error term where everything is expected to deviate at random from the expected value, this only represents the random deviations for each school. s_i is distributed $s_i \sim N(0, \sigma_s^2)$ where σ_s is the standard deviation of differences of each school's win percentage relative to its predicted value based on APR scores and other fixed effects. It is also important to add a similar random effect, p_{ij}, that allows for deviations between programs at the same school after accounting for APR and other fixed effects. p_{ij} is distributed $p_{ij} \sim N(0, \sigma_p^2)$ where σ_p is the standard deviation of differences of each specific program's win percentage relative to its predicted value based on APR scores and other fixed effects. As in the OLS model, the mixed effect models still have an error term ϵ_{ijk}, however now this random effect represents the unexplained variability in win percentage within a program from year-to-year.

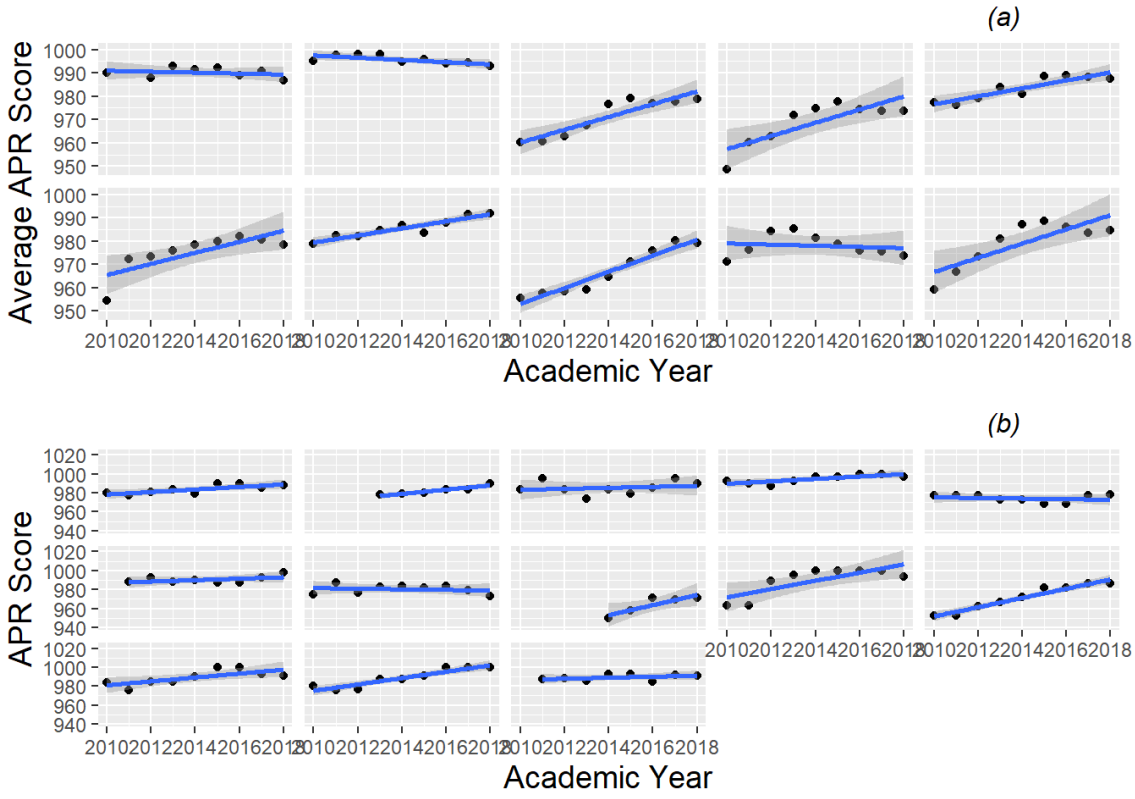

Figure 4: Differences in slopes (a) from school to school selecting at random 10 schools, and (b) from program to program selecting all programs from 2 schools at random

We have considered random effects allowing for random deviations for schools, programs, and years, but the inclusion of another random effect is important: one associated with differences in rates of change from school-to-school or program-to-program. In figure 4 (above) we can see a sample of schools (a) and how their average APR score has a large variation in slopes, some being moderately positive and some being slightly negative. In (b) we can see from a sample of programs from 2 different schools that programs can have slight variations in slope as well. In our models this will be the m_i and m_{ij} variables. Both of these are again normally distributed with mean 0 and variance σ_m^2.

Model 4: $Y_{ijk} = \alpha_0 + \beta_1 APR_{ijk} + s_i + p_{ij} + \epsilon_{ijk}$

Model 5: $Y_{ijk} = \alpha_0 + \beta_1 APR_{ijk} + s_i + m_i APR + p_{ij} + \epsilon_{ijk}$

Model 6: $Y_{ijk} = \alpha_0 + \beta_1 APR_{ijk} + s_i + p_{ij} + m_{ij} APR + \epsilon_{ijk}$

When we add these random effects to the fixed effects in Model 1, we have 3 different base models. Model 4 includes only intercept random effects, while Model 5 and 6 include slope random effects for schools and programs respectively. In other words, this allows for the rate of change in Win Percentage to differ based on the school or program's average APR score. A school or program with a lower APR score may see a sharper increase in Win Percentage than schools or programs with a higher APR score. We can use likelihood ratio tests to compare our models. In doing so, the models are refitted using full maximum likelihood, since we are focusing on model parameters and not random effects. We can see the comparison in Table 3 (below). In this case, the AIC favors Model 6 (145425) over Model 5 (145538) and Model 4 (145615). The BIC favors Model 6 as well (145480) over Model 5 (145593) and Model 4 (145654). We can conclude that Model 6 outperforms both Model 4 and Model 5.

Table 3: Likelihood Ratio Test using anova

	npar	AIC	BIC	logLik	dev	Chisq	Df	pval
Model 4	5	145615.3	145654.2	-72802.64	145605.3	NA	NA	NA
Model 5	7	145538.4	145592.9	-72762.21	145524.4	80.86436	2	0
Model 6	7	145425.2	145479.7	-72705.62	145411.2	113.18540	0	NA

Model 6.a:

$$Y_{ijk} = \alpha_0 + \beta_1 APR_{ijk} + \beta_2 Enrollment_i + \beta_3 PercentWhite_i + s_i + p_{ij} + m_{ij} APR + \epsilon_{ijk}$$

Model 6.b:

$$\begin{aligned} Y_{ijk} =& \alpha_0 + \beta_1 APR_{ijk} + \beta_2 Enrollment_i + \beta_3 PercentWhite_i \\ &+ \beta_4 PercentMale_i + \beta_5 PercentCommunication_i + \beta_6 NewEngland_i \\ &+ \beta_7 MidEast_i + \beta_8 GreatLakes_i + \beta_9 Plains_i + \beta_{10} SouthEast_i \\ &+ \beta_{11} SouthWest_i + \beta_{12} RockyMountains_i + \beta_{13} FarWest_i \\ &+ \beta_{14} Outlying_i + s_i + p_{ij} + m_{ij} APR + \epsilon_{ijk} \end{aligned}$$

We can also further analyze model 6 to see if it can be improved by adding more fixed effects, other than APR Score, from the College Scorecard Database. Similar to models 2 and 3, the second equation uses APR in addition to the most correlated variables from the College Scorecard Database and Model 3 depicts a model that also includes the variables that are very slightly correlated in addition to dummy variables for the different geographic regions.

$$\text{Pseudo } R^2_{L3_s} = \frac{\hat{\sigma_s}^2(\text{Model 6}) - \hat{\sigma_s}^2(\text{Model 6.a})}{\hat{\sigma_s}^2(\text{Model 6})} = \frac{41.21 - 24.95}{41.21} = 0.3946$$

$$\text{Pseudo } R^2_{L3_s} = \frac{\hat{\sigma_s}^2(\text{Model 6.a}) - \hat{\sigma_s}^2(\text{Model 6.b})}{\hat{\sigma_s}^2(\text{Model 6.a})} = \frac{24.95 - 21.10}{24.95} = 0.1543$$

In an ordinary $\text{Pseudo } R^2$ calculation, we would be looking at σ^2 to determine the impact of the different models on year to year variability within programs, however no variables are added at the yearly, or program level, so when adding College Scorecard variables to Model 6, there will be no change in program slope or intercept terms. To analyze the addition of these variables to the model we can look at the $\text{Pseudo } R^2_{L3_s}$ using σ^2_s to describe the improvement from model to model in explaining school-to-school variability in intercepts. The addition of these variables at the school level has decreased the between-school variability in win percentage by 39.46% from model 6 to 6.a and an additional 15.43% from model 6.a to 6.b. Using AIC and BIC tests in Table 4 (below), we see that AIC favors model 6.b, but BIC favors model 6.a due to the fact that BIC has a larger penalty for each additional variable and more model complexity. Due to the improvement in models using $\text{Pseudo } R^2_{L3_s}$ and the slight favoritism of Model 6.b with AIC, Model 6.b is the final model that will be used throughout the remainder of our research.

Table 4: AIC and BIC Comparisons

	df	AIC	BIC

	df	AIC	BIC
Model 6	7	145432.6	145487.0
`Model 6.a`	9	145162.5	145232.5
`Model 6.b`	19	145123.7	145271.6

Results

As stated in the methodology section, Model 6.b (shown below) is the model that will be analyzed to draw our results. This three-level mixed effects model includes fixed effects for APR Score and all the variables that are slightly correlated, dummy variables for the different geographic regions, and random effects for school-to-school variability, program-to-program variability, changes in rates of change from program to program, and an error term for unexplained year-to-year variability.

Model 6.b:

$$\begin{aligned} Y_{ijk} =& \alpha_0 + \beta_1 APR_{ijk} + \beta_2 Enrollment_i + \beta_3 PercentWhite_i \\ &+ \beta_4 PercentMale_i + \beta_5 PercentCommunication_i + \beta_6 NewEngland_i \\ &+ \beta_7 MidEast_i + \beta_8 GreatLakes_i + \beta_9 Plains_i + \beta_{10} SouthEast_i \\ &+ \beta_{11} SouthWest_i + \beta_{12} RockyMountains_i + \beta_{13} FarWest_i \\ &+ \beta_{14} Outlying_i + s_i + p_{ij} + m_{ij} APR + \epsilon_{ijk} \end{aligned}$$

Table 5: Mixed Effects Model 6.b Results

	Win Percentage		
Predictors	*Estimates*	*SE*	*T Stat*
Intercept	38.572	4.895	7.879
Centered APR Score	0.104	0.010	10.544
Total Enrollment	0.000	0.000	7.905
Percent White	0.118	0.016	7.494
Percent Male	0.011	0.043	0.263
Percent Communication Majors	0.226	0.070	3.213
New England	-3.744	4.025	-0.930
Mid East	-3.618	3.933	-0.920
Great Lakes	-3.528	3.977	-0.887
Plains	-0.444	4.095	-0.108
Southeast	0.560	3.931	0.142

Southwest	1.978	4.071	0.486
Rocky Mountains	-1.894	4.187	-0.452
Far West	1.626	4.030	0.403

Random Effects

σ^2	167.14631
$\tau_{00\ Program}$	85.67472
$\tau_{00\ School}$	21.10242
$\tau_{11\ Program.centeredAPR}$	0.05187
$\rho_{01\ Program}$	0.14454
$N_{\ School}$	374
$N_{\ Program}$	2168
Observations	17661

From the model results shown in Table 5, a total 17,661 observations from 374 different schools and 2,168 different programs were used to estimate our model. We can begin to interpret our results by looking at the fixed effects estimates. $\alpha_0 = 38.572$ and is the mean win percentage of an athletic program with a mean APR Score of 974.4 at a U.S. service school where enrollment is 0 students, 0 percent of which are white, 0 percent of which are male, and 0 percent are communication majors. Of course there is no school in our data where this holds, so this estimate is not a valid interpretation of any school in our dataset. $\beta_1 = 0.104$. This is the estimated increase in a program's win percentage for a 1 point increase in APR score after controlling for total enrollment, the percent white, the percent male, the percent communications majors, and the geographic region. With a T-Statistic of 10.544, this is statistically significant and is the variable of interest in our research. $\beta_2 = 0.000267$ and corresponds to the estimated increase in win percentage for one more enrolled student at the school after controlling for APR score, the percent white, the percent male, the percent communications majors, and the geographic region. A single student increase is not reasonable, so we can say that a 1,000 student increase in enrollment corresponds to 0.267 percentage point increase in win percentage. With a T-Statistic of 7.905, this is also statistically significant. We can see $\beta_3 = 0.118$ which is the estimated increase in win percentage for a 1 percentage point increase in the percentage of the student body that is white after controlling for APR score, total enrollment, the percent male, the percent communications majors, and the geographic region. The T-Statistic of 7.494 shows that this is significant. $\beta_4 = 0.011$ and is the estimated increase in win percentage for a 1 percentage point increase in the percentage of the student body that is male after controlling for APR score, total enrollment, the percent white, the percent communications majors, and the geographic region. The small T-Statistic of 0.263 shows this is insignificant. $\beta_5 = 0.226$. This is the estimated increase in win percentage for a 1 percentage point increase in the percentage of the student body that is majoring in communications or a related field after controlling for APR score, total enrollment, the percent white, the percent male, and the geographic region. The T-Statistic of 3.213 is significant, but not as much as APR Score, Enrollment, or Percent White.

We can continue our results interpretation by looking at the variables corresponding to the difference in win percentage for different geographic regions, compared to U.S. Service Schools. The low T-Statistics demonstrate that these variables are statistically insignificant. $\beta_6 = -3.744$ and is the estimated decrease in win percentage for schools that are in the states of Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont. $\beta_7 = -3.618$, which is the estimated decrease in win percentage for schools that are in the states of Delaware, District of Columbia, Maryland, New Jersey, New York, and Pennsylvania. $\beta_8 = -3.528$; this corresponds to the estimated decrease in win percentage for schools that are in the states of Illinois, Indiana, Michigan, Ohio, and Wisconsin. $\beta_9 = -0.444$ and is the estimated decrease in win percentage for schools that

are in the states of Iowa, Kansas, Minnesota, Missouri, Nevada, North Dakota, and South Dakota. $\beta_{10} = 0.560$ and is equal to the estimated decrease in win percentage for schools that are in the states of Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia. $\beta_{11} = 1.978$; the estimated decrease in win percentage for schools that are in the states of Arizona, New Mexico, Oklahoma, and Texas.$\beta_{12} = -1.894$ and equals the estimated decrease in win percentage for schools that are in the states of Colorado, Idaho, Montana, Utah, and Wyoming. $\beta_{13} = -1.626$; the estimated decrease in win percentage for schools that are in the states of Alaska, California, Hawaii, Nevada, Oregon, and Washington.

It is also important to interpret the random effects estimates we see in Table 5. $\sigma = \sqrt{1}67.146 = 12.929$. This is the standard deviation in within-program residuals from year-to-year after accounting for APR score, total enrollment, the percent white, the percent male, the percent communications majors, and the geographic region. $\sigma_p = \sqrt{8}5.675 = 9.256$ and corresponds to the standard deviation of differences of each program's win percentage at a specific school relative to its expected win percentages based on APR score, total enrollment, the percent white, the percent male, the percent communications majors, and the geographic region. $\sigma_s = \sqrt{2}1.102 = 4.954$; the standard deviation of differences of each school's win percentage relative to its expected win percentages based on APR score, total enrollment, the percent white, the percent male, the percent communications majors, and the geographic region. $\sigma_m = \sqrt{0}.052 = 0.228$ and is equal to the standard deviation in rates of change in win percentage relative to APR scores. $\rho_{01} = 0.145$ and equals the correlation in programs' random intercept and their random rates of change. This means that programs with an above average APR score see sharper increases in Win Percentage on average, however this is a small correlation.

Based on T-values produced by Model 6.b, APR scores have the most significant effect on a program's win percentage. As stated above, on average a 1 point increase in APR score corresponds to a 0.104 percentage point increase in win percentage after controlling for all other factors. However, a 1 point increase is not a reasonable scale when the range of APR scores in our dataset are from 451 to 1,000. Instead we can interpret this by saying a 10 point increase in APR score is associated with a 1.04 percentage point increase in a program's win percentage, and a 100 point increase corresponds to a 10.4 percentage point increase in win percentage. A 100 point difference in APR score is therefore equivalent to roughly winning 1 more game out of every 10 games. From our results and the corresponding T-Statistics, the estimates for APR Score, Total Enrollment, Percent White, and Percent Communication Majors are statistically significant. Percent male along with the differences between geographic regions are statistically insignificant; they do not significantly differ from 0.

Controlling for school level variables provided more reliable estimates of the effects of APR scores, while also providing interesting insights. For example, schools with larger enrollment and schools with a larger percent of the student body being white had higher win percentages on average. In addition, the inclusion of random effects in our model makes our variance and standard deviation estimates more precise. However, it is still important to note that the largest amount of deviation occurs from year-to-year within programs rather than from school-to-school or program-to-program within school.

Conclusion

Transitioning from the OLS model and comparing various mixed effects models using AIC, BIC, and likelihood ratio tests allow our final model to yield the most accurate results. Using our final mixed effects model of 2,168 different programs over eight years, we see the direct positive relationship between academic success, measured through APR scores, and program's win percentage, a measure of athletic performance. Of the numerous covariate factors included in this model, academics have the most significant effect on athletics. The NCAA and all college coaches and recruiters for team sports should be looking not just for the best athletes, but for the best student-athletes. In

addition, the NCAA and colleges should consider this as an incentive for programs to encourage and develop higher academic performance, as it will benefit the competition at the college-level and create better students, better athletes, and in general people who are better prepared for their life after college.

This significant finding is intuitively the results we expected to find in our research. The mentality of hard-working students is directly reflected in their athletic preparation and performance. Future research on the relationship between academics and athletics could look at the relationship for non-team related sports, such as track and field or swimming. This relationship could also be viewed at different levels of competition, perhaps at the high school, or NCAA Division 2 or Division 3 level.

References

"Age, Gender, Ethnicity, Religion and Culture and Sport - Social Groupings and Participation in Sport - OCR - GCSE Physical Education Revision - OCR - BBC Bitesize." BBC News, BBC, https://www.bbc.co.uk/bitesize/guides/zy62hv4/revision/2 (https://www.bbc.co.uk/bitesize/guides/zy62hv4/revision/2).

Ayers, Kevin, Monica Pazmino-Cevallos, and Cody Dobose. "The 20-hour rule: Student-athletes time commitment to athletics and academics." Vahperd Journal 33.1 (2012): 22-27.

Battle, Johné Everett. The Talent Factory: A Study of Grit, Mindset, and Student-Athletes. Diss. University of Pennsylvania, 2020

Gurney, Gerald S. "Now we must reform athletics reform." The Chronicle of Higher Education 56.9 (2009): A34

Lawrence, Janet, Molly Ott, and Lori Hendricks. "Athletics reform and faculty perceptions." (2009).

Legler, Julie, and Paul Roback. "Beyond Multiple Linear Regression." Beyond MLR, 26 Jan. 2021, https://bookdown.org/roback/bookdown-BeyondMLR/ (https://bookdown.org/roback/bookdown-BeyondMLR/).

McCormick, Kaydee K. "Academic Clustering in Intercollegiate Athletics." Kansas State University, 1 Jan. 1970, https://krex.k-state.edu/dspace/handle/2097/4129 (https://krex.k-state.edu/dspace/handle/2097/4129).

Oskolkov, Nikolay. "Maximum Likelihood (ML) vs. REML." Medium, Towards Data Science, 9 Sept. 2020, https://towardsdatascience.com/maximum-likelihood-ml-vs-reml-78cf79bef2cf (https://towardsdatascience.com/maximum-likelihood-ml-vs-reml-78cf79bef2cf).

"PROFESSIONAL ATHLETE Demographics And Statistics In The US." Zippia, https://www.zippia.com/professional-athlete-jobs/demographics/ (https://www.zippia.com/professional-athlete-jobs/demographics/).

Schneider, Ray G., et al. "Academic Clustering and Major Selection of Intercollegiate Student-Athletes." College Student Journal, vol. 44, no. 1, Mar. 2010, pp. 64–70. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&db=aqh&AN=48646428&site=eds-live.

Song, Minyong, and Yang Zhang. "Research on the Relationship between Geographical Factors, Sports and Culture." Advances in Physical Education, vol. 08, no. 01, Feb. 2018, pp. 66–70., https://doi.org/10.4236/ape.2018.81008 (https://doi.org/10.4236/ape.2018.81008).

Data Sources

Academic Progress Rate, NCAA, https://web3.ncaa.org/aprsearch/aprsearch (https://web3.ncaa.org/aprsearch/aprsearch).

"Data Documentation." Documentation | College Scorecard, https://collegescorecard.ed.gov/data/documentation/ (https://collegescorecard.ed.gov/data/documentation/).

NCAA Statistics, NCAA, https://stats.ncaa.org/rankings/change_sport_year_div (https://stats.ncaa.org/rankings/change_sport_year_div).

Appendix

Table 6: OLS Results

	Dependent variable:		
	Win Percentage		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
Centered APR Score	0.119***	0.090***	0.105***
	(0.006)	(0.006)	(0.006)
Total Enrollment		0.0003***	0.0003***
		(0.00001)	(0.00001)
Percent White		0.090***	0.105***
		(0.006)	(0.007)
Percent Male			0.024
			(0.019)
Percent Communication Majors			0.255***
			(0.033)
New England			-3.445**
			(1.492)
Mid East			-3.434**
			(1.466)
Great Lakes			-3.497**
			(1.487)
Plains			0.013
			(1.541)
Southeast			0.718
			(1.475)
Southwest			2.101
			(1.532)
Rocky Mountains			-1.818
			(1.588)
Far West			1.258
			(1.510)
Constant	50.089***	40.089***	38.303***
	(0.132)	(0.426)	(1.957)
Observations	17,685	17,661	17,661
R^2	0.023	0.073	0.089
Adjusted R^2	0.023	0.073	0.088
Residual Std. Error	17.533 (df = 17683)	17.079 (df = 17657)	16.941 (df = 17647)
F Statistic	423.779*** (df = 1; 17683)	466.037*** (df = 3; 17657)	132.306*** (df = 13; 17647)
Note:			*p<0.1;* ***p<0.05;*** p<0.01