Football and Enrollment Correlations

ABSTRACT

In this project I seek to find whether changes in college football teams' seasonal win records prompt a noticeable change in applications to those universities in the following year. For each major college football conference I generate a scatterplot and linear regression of its data points. For each individual team I generate a density plot of bootstrapped r-squared values and also create a scatterplot. Finally, I summarize my findings.

This project began after an informal observation of enrollment changes at Kansas State University (K-State), the local university in my hometown. Between 2010 and 2012 K-State's football team had phenomenal success - the team was ranked #1 in the nation for a time. During this same time period the university experienced record student enrollment rates. Meanwhile, K-State's in-state rival, the University of Kansas (KU), experienced some of the worst football seasons in university history during this timeframe. The university also suffered declining enrollment. Only a few years later, though, KSU began experiencing declining enrollment as its football team regressed. This made me wonder: do most large universities experience surges in enrollment when their football teams perform well and declines in enrollment when their teams struggle?

To answer my question I decided to analyze university application data and football team records (wins vs losses) for a period of approximately 10 years. I gathered enrollment data from the National Center for Education Statistics and football data from the NCAA. I looked only at Power 5 universities - that is, universities whose football teams are members of either the Big 10, Big 12, SEC, ACC, or Pac 12 athletic conferences. I reason that these schools recruit students nationally rather than regionally, so their student bodies are more likely to change based on the performance of their football teams than are the student bodies of smaller, regional colleges. I chose not to look at data for universities that changed athletic conferences during the time period of interest since that change could effect enrollment changes.

There are two variables of interest in my model: percent change in football wins and percent change in college applications. Looking at percent change is more insightful than looking at raw changes in wins and applications. For example, it is a greater change to go from winning two games to four in a twelve-game season than it is to go from winning seven games to nine, and fans probably prefer large improvements to small ones. It is possible that the number of games won is more important than the change in the percentage won - perhaps when a college suddenly wins enough games to enter the AP list of top 25 programs or become eligible to play in a bowl game it causes a spike in applications - but that is an assumption for a different model.

Results are shown in a scatterplot since this type of graph allows the values of both variables to be displayed for many different data points. To plot each data point the change in applications for each year is matched with the change in football wins that occurred in the previous year. For example, a datapoint representing the change in applications between 2010 and 2011 will be matched with football data representing the change in wins from 2009 to 2010. The reasoning for modeling the data in this manner is that applicants apply before seeing the end of the current season and therefore rely on data from the previous season. Because both application and football seasons begin at roughly the end of August it is possible that applicants are influenced by more recent data when they make application decisions, but it is assumed that this is not the case.

A density plot of R-squared values is shown to provide context for how unusual each college's observed R-squared value may be. The plot uses 1000 bootstrapped samples taken from eleven years of observed data for each college. A necessary assumption is that eleven data points are not too few to bootstrap.

My model shows a weak connection between football wins and college applications. The slope of the regression line relating change in football wins to change in college applications is nearly flat for the data as a whole. When results are narrowed by conference the slope of the resulting regression line is marginally steeper for some conferences (Big 10, Big 12, Pac 12) but has a slight negative slope for others (ACC).

The data becomes more interesting when analyzed by college. Some colleges seem to be affected by outliers - for example, Georgia Tech's r-squared estimate of percent change in enrollment per percent change in football wins appears to be roughly 0.125, a rather low value. However, a 95% confidence interval created from bootstrapping 1000 samples from the dataset contains r-squared values up to approximately r-squared = 0.625. Plotting a scatterplot of data points makes it evident that there is one point on the far left that is skewing results. Some universities have r-squared values that are relatively high. Purdue, for instance, has an r-squared value approximately equal to 0.45, while Washington State has a r-squared value of 0.625. An overwhelming majority of universities, though, have low r-squared values and scatterplots with points that are fairly spread out. Some colleges, like Iowa State, have regression lines with negative slopes as a result.

Figure 1: Washington State has the highest R-squared value of all colleges studied

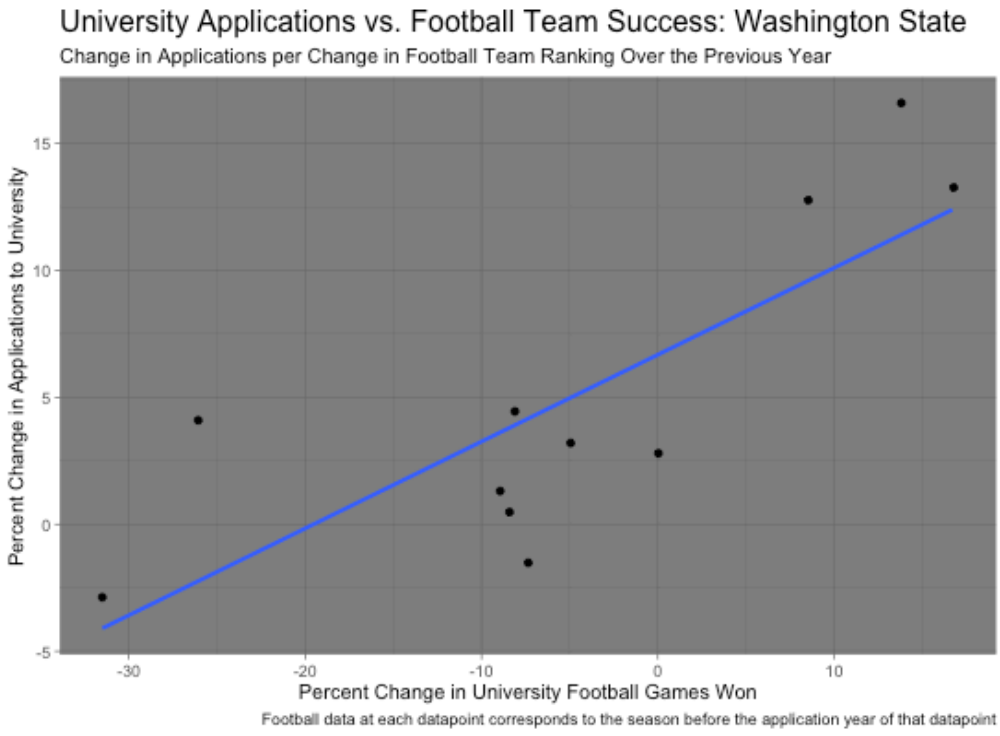

Figure 2 - Iowa State's regression line has a negative slope, but Iowa State has a low R-squared value and its points are widely spread out.