Diego Valdes

IST 718

Lab 1

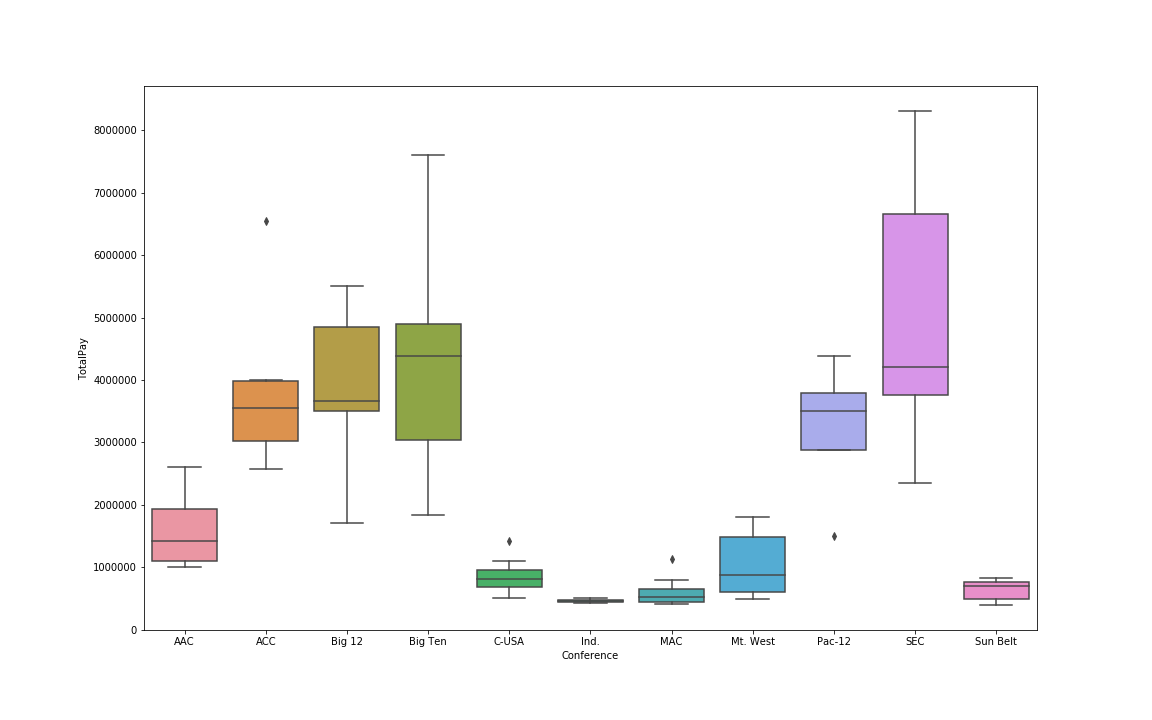
July 19, 2019

**Data Exploration**

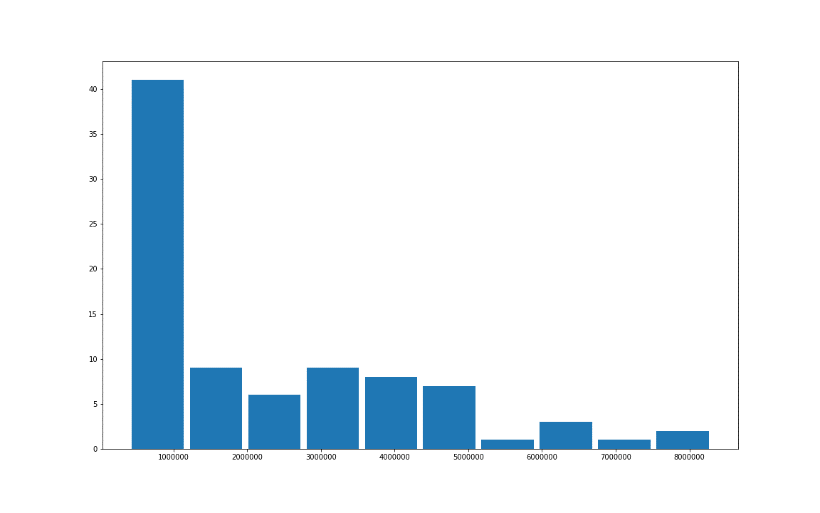
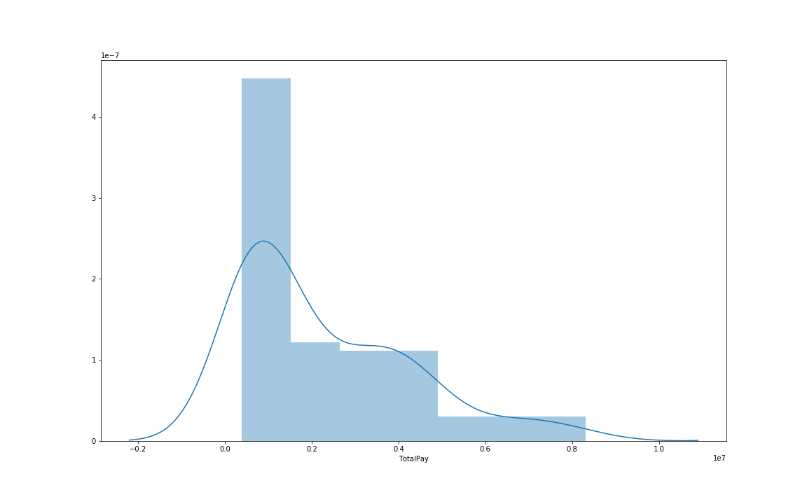
The analysis presented in this paper was developed using data obtained both from course work and the NCAA site. The course work data is comprised of coaches salaries at universities at 11 conferences (one of those ‘conferences’ for the purpose of analysis is independent). It has their base pay, bonuses, assistant pay, the university they coach for, and the conference they play in. The NCAA data is comprised of many attributes, but for the purpose of this analysis, only graduation rates (GSR, FED\_RATE) were used. These two sets were merged with the key field ‘Conference’ using the mean for the graduation rates based on conference.

There were no NaN in either dataset, but in the coaches data there were instances of fields being populated by ‘—‘. All rows that had these types of entries were dropped. This brought the coaches data from 129 rows to 87.

The box plot below shows the biggest range of salaries can be found in the SEC and Big Ten, with many of the higher salaries belonging to the SEC. Many football fans wouldn’t find this surprising, as Alabama, recent multiple NCAA champions reside in the SEC. Most informative is the lowest ranges fall within the Independent schools, which shows why the most advantageous jobs fall within the conferences.



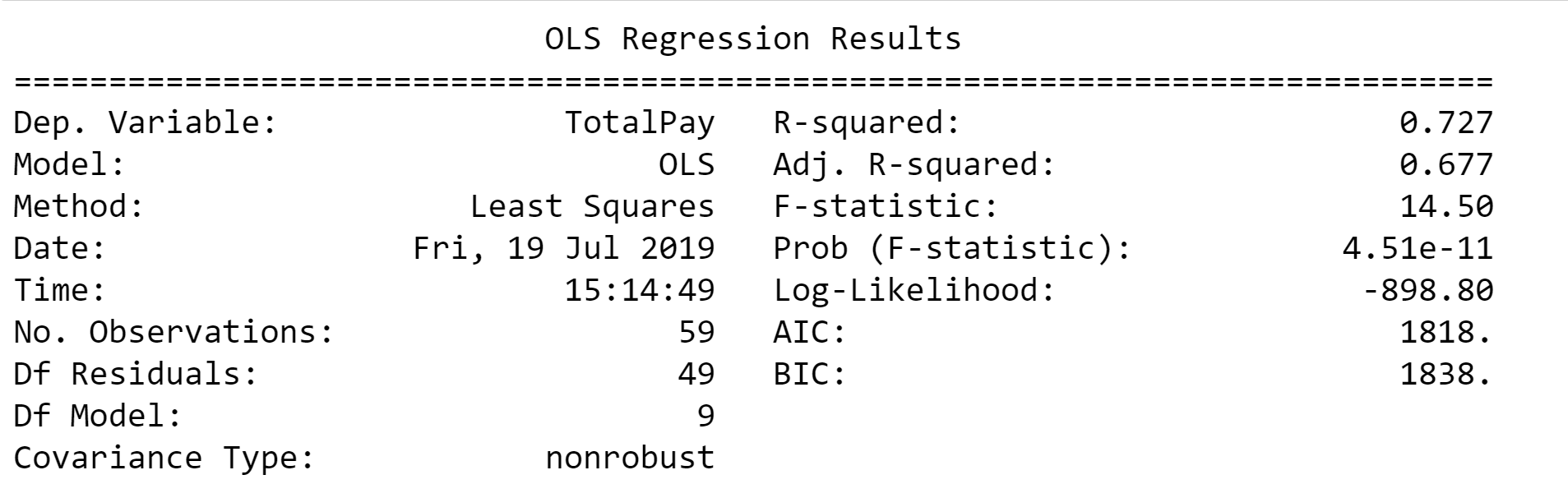
The histograms below show the distribution of pay. Most coaches salaries do come closer to the $1,000,000 price tag.



**Models**

Multiple regression models where built using various attributes. Attributes where removed based on significance in test models. After multiple attempts, there was no significant difference between the type of regression models (ols, wls, gls), and removal of the initial variables (conference, bonus paid, assistant pay, buyout, fed rate, gsr) did not produce better results, even when the significance of the conference variable was poor, ie: Conference[T.Atlantic Coast Conference] at .345 vs Conference[T.Southeastern Conference] at .001. This observation, which occurred for multiple conference variables, demonstrates that a conference may have an impact on pay, but not necessarily. Clearly, a coach’s salary in the SEC is more impacted by being in that conference than a coach in the ACC.

Based on these tests, the ols regression model was used moving forward. Although the monetary variables showed a significant impact on total pay, which is an obvious observation not necessary of a regression model, they were removed to better assess the impact of the conference and graduation rates. These models weren’t as strong as those before (R-squared of .93 vs .72), but the results were within a tolerable acceptance rate.



This model used the attributes conference, fed rate, and gsr. The graduation rates remained significant in the model and the conference attribute continued to demonstrate the same patterns of significance/insignificant as with previous models as discussed earlier.

**Results**

**What is the recommended salary for the Syracuse football coach?**

Using the OLS model, the recommendation for total compensation for the Syracuse football coach would be $2,543,443.

**What would his/her salary be if the university was still in the Big East?**

Unfortunately, the coaches data set did not include the salaries of the Big East. But a simple Google query provided this data via ESPN’s website. Due to the time constraints of generating this analysis, this data was not included in the modeling and the best estimate for this question is the mean, $1,401,142.

**What would his/her salary be if the university was still in the Big East?**

Using the OLS model, the recommendation for total compensation for the Syracuse football coach would be $3,999,111.

**What schools were dropped the data and why?**

Schools that had incomplete data where dropped from the coaches dataset. The NCAA data had more schools and conferences, and since there weren’t any matching data to join to, these entries were not included. Of the 32 conferences included, only 10 of them where used.

**What effect does graduation rate have on projected salary?**

Both graduation variables used were significant in each regression model, and shared the similar characteristic that the fed rate was positively correlated and the gsr was negatively correlated. In the model used to predict salaries, the gsr was more negatively impactful than in previous models (~ -2.4e+5 vs ~ -8.4e+4). The gsr is more closely tied in with graduation rates of student athletes, which goes to demonstrate that while the graduation rate does have an effect on the coaches salary, it is not a factor in making the ridiculous salary that many coaches receive.

**Areas of Opportunity for Further Analysis**

Although this analysis is adequate it is far from close to being a complete product. Better models could be generated using more data, specifically, win/loss (overall and conference) records of the schools, number of championships, players sent to the NFL (by round if available), and number of high desired high school recruits. This data was not included, not because it doesn’t exist, but because the writer misjudged the due date of the assignment and was forced to put it together with less than the expected time. This lack of planning, in combination with the discoveries made during the exploration process revealed that much more data would be needed to better provide a prediction for salaries.