Jose Reyes
7.23.2020
IST 718

Lab 3

Objective

The salary of the next Syracuse football coach is determined by creating linear regression models based on various input variables. Several datasets that include football records, stadium sizes, graduation rates and offensive and defensive stats across all Division 1 College Football Programs were mined and merged together into one data frame. These variables are first analysed to determine if they have a relationship with salary size. They are then used as inputs into two linear regression models which are compared for accuracy and used to predict the next Syracuse coach's salary.

Load and clean Datasets

The following datasets were loaded and cleansed in Python using the Pandas package:

- **Coaches9.csv** is the initial dataset provided which contains information on 130 schools including the athletic conference and coach salary.
- **CollegeRecords.csv** Includes each school's 2019-2020 win and loss record and was extracted from https://collegefootballdata.com/exporter/records.
- **Stadiums.csv** Includes each school's stadium capacity and was extracted from https://www.collegegridirons.com/comparisons-by-capacity/.
- **2006 Graduation Rates.csv** includes the graduation success rate for each school and was extracted from http://www.ncaa.org/about/resources/research/graduation-rates. Per the project instructions, the 2006 cohort was used and includes both GSR and FGR.
- **Offense_Defense.csv** Includes each school's offensive and defensive stats for the 2019-2020 year including points per game (PPG) and the average number of points allowed (Avg). Data was extracted from https://www.ncaa.com/stats/football/fbs/current/team/28.

All datasets were merged together into one data frame for analysis. The merge was completed by ensuring each data set has matching school names.

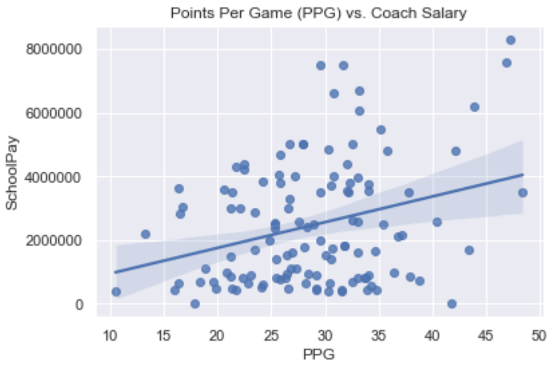
Data from the following schools were dropped due to missing data:

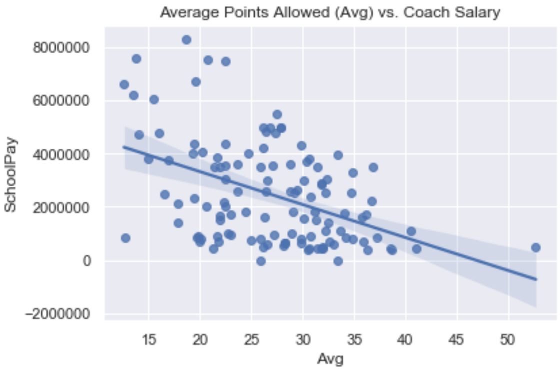
- Missing Stadium Size: Liberty
- Missing Coach Salary: Baylor, Brigham Young, Rice, and Southern Methodist
- Missing 2006 Graduation Rates: Charlotte, South Alabama, Texas-San Antonio, Georgia State

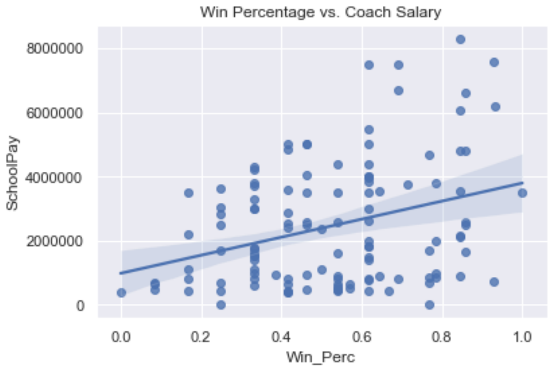
Jose Reyes
7.23.2020
IST 718

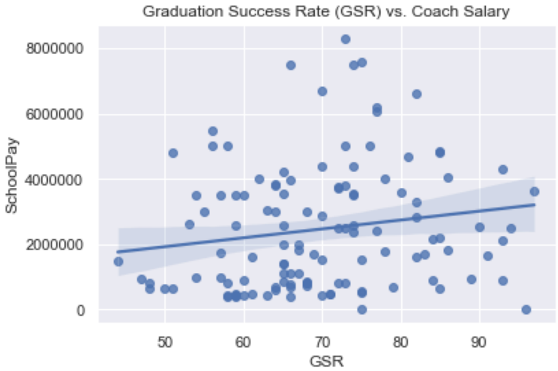
Analysis

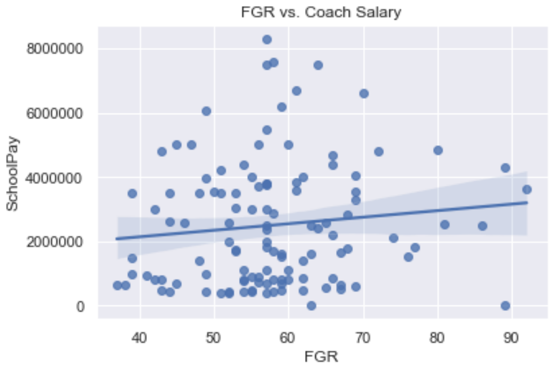
The following scatterplots evaluate several variables' relationships with salary size:

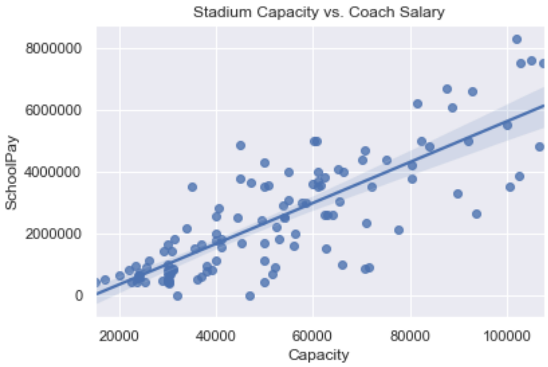








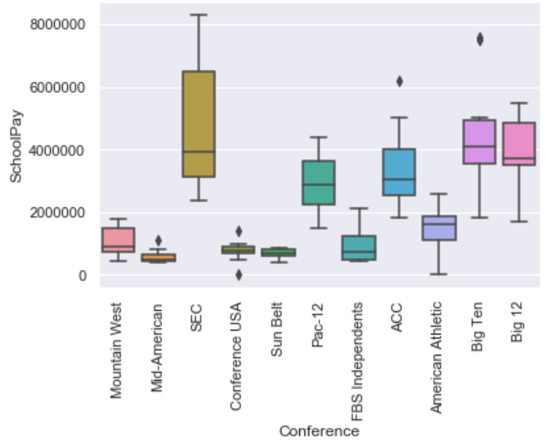


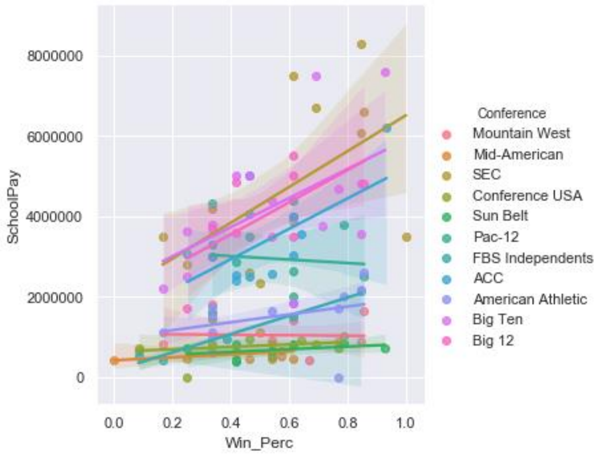


Unsurprisingly, as a team is more successful by winning more games, scoring higher, and allowing less points, a coach's salary is generally higher. Similarly, the stadium size tends to hold a higher capacity as successful teams tend to draw a higher audience. The graduation rates, indicated by GSR and FGR, have a slightly positive relationship but is much less pronounced compared to other variables. Intuitively, graduation rates are harder to link to a coach's salary as this is not part of their job.

The following two graphs assess the relationship between the school's athletic conference and salary.

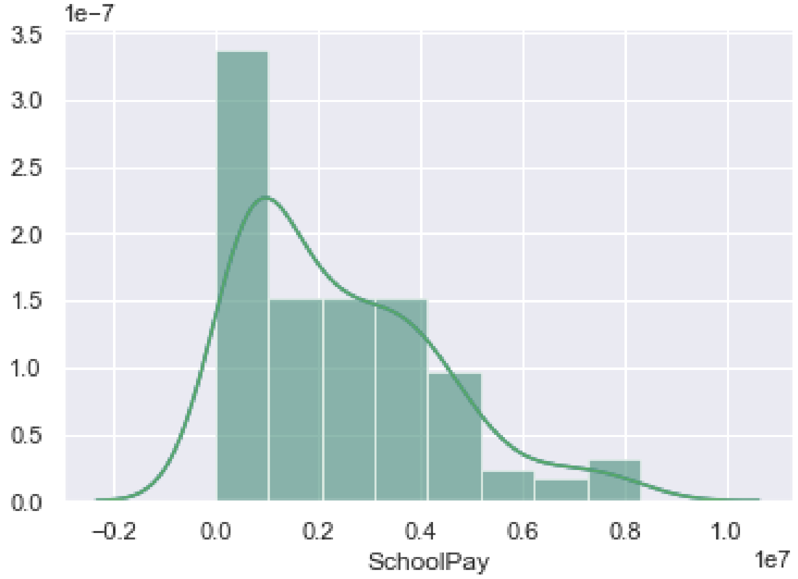
Jose Reyes
7.23.2020
IST 718





Salaries in the Power 5 (P5) conferences (SEC, Big10, Big12, ACC, and Pac-12) are clearly the highest with SEC having the highest of all conferences. Mountain West, Mid-American, Conference USA, and Sun Belt have similarly smaller salaries. Additionally, salaries tend to increase by win percentage and the most successful teams in the SEC have some of the highest salaries in all of college football.

We can now take a preliminary look at how Syracuse stacks up against all other schools.



This graph shows that the distribution of salaries across all Division 1 Football Programs is positively skewed with a longer right tail. Most salaries are concentrated on the left with a majority being below $2 million. The Syracuse coach’s salary is currently $2,401,206.

Syracuse’s stats are compared with the average stats across all other Division 1 football programs in this table.

Jose Reyes
7.23.2020
IST 718

	PPG	AVG	Win_Perc	GSR	Capacity
Division 1 Football Programs Mean	28.73	26.97	0.52	69.62	52,130.40
Division 1 Football Programs Standard Deviation	7.082	6.90	0.22	11.55	23,480.79
Syracuse	**28.30**	**30.70**	**0.42**	**77**	**49,250.00**

Overall, Syracuse performs lower than average in almost all variables except for GSR. However, Syracuse is in the ACC which has higher salaries than all other schools in non-P5 conferences as well as the Pac-12. It will be determined in the modelling how each of these variables will contribute to the salary size and by how much.

Modelling

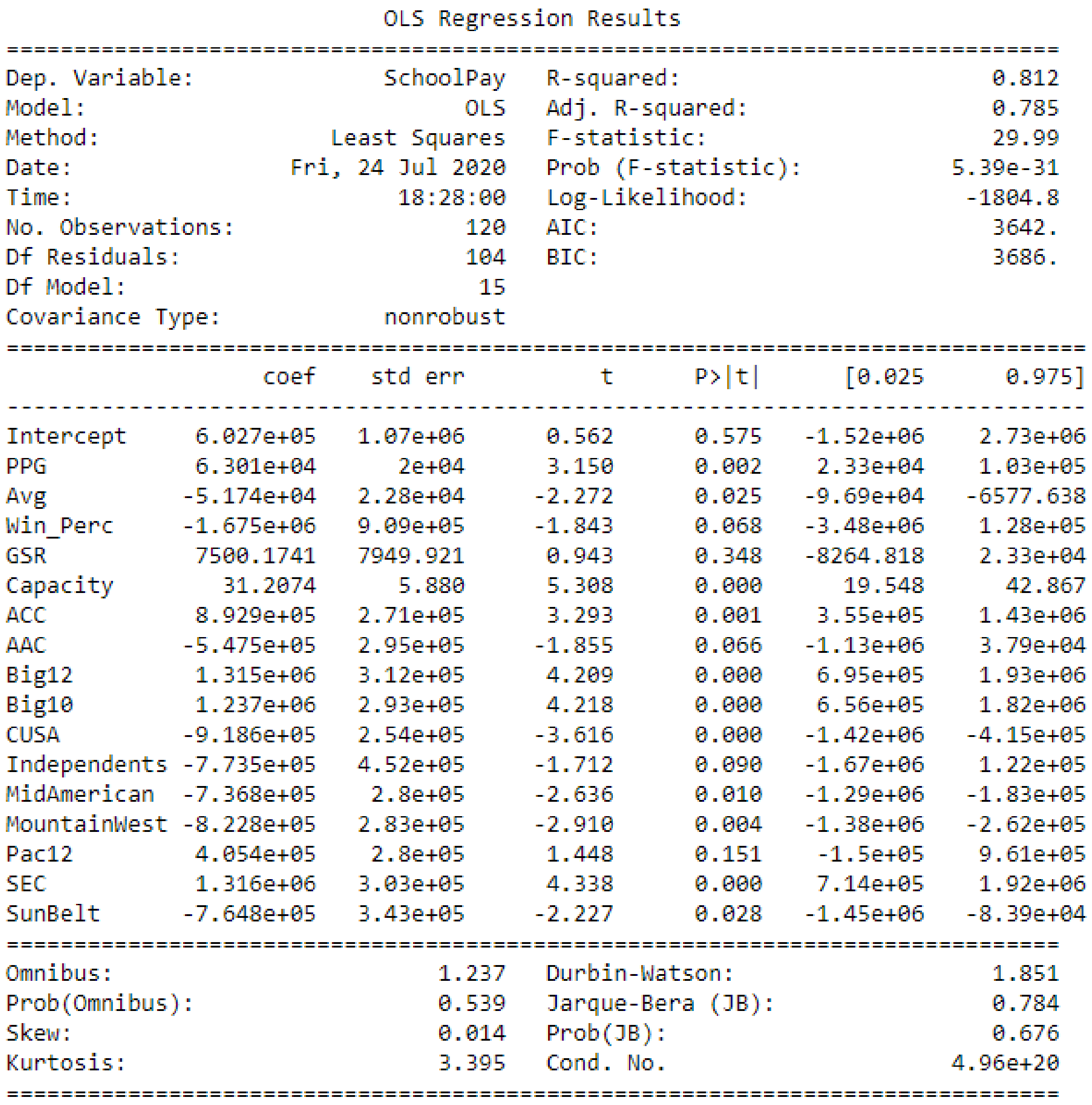
Based on the exploratory analysis, the following variables were used as inputs for the linear modeling:

- Points Per Game (PPG)
- Average Points Allowed (Avg)
- Win Percentage (Win_Perc)
- GSR (Graduation Success Rate)
- Capacity
- Conference

The FGR variable was dropped to avoid multicollinearity as it had an almost identical relationship with coach salary as GSR. Because conference is a nonordered categorical variable, it is converted into dummy variables.

Jose Reyes
7.23.2020
IST 718

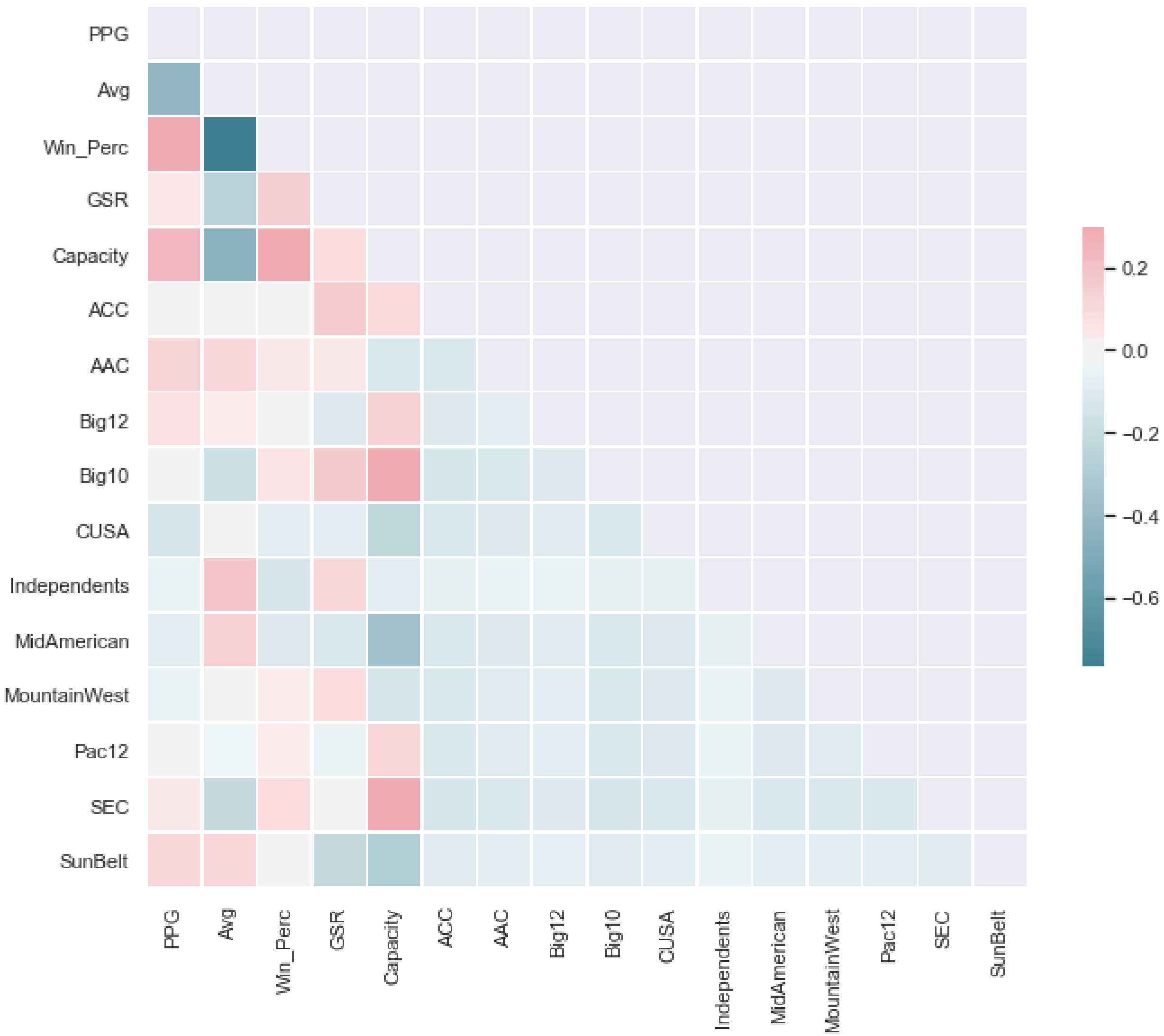
Model 1 results:



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              SchoolPay   R-squared:                       0.812
Model:                            OLS   Adj. R-squared:                  0.785
Method:                 Least Squares   F-statistic:                     29.99
Date:                Fri, 24 Jul 2020   Prob (F-statistic):           5.39e-31
Time:                        18:28:00   Log-Likelihood:                -1804.8
No. Observations:                 120   AIC:                             3642.
Df Residuals:                     104   BIC:                             3686.
Df Model:                          15                                         
Covariance Type:            nonrobust                                         
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept     6.027e+05   1.07e+06      0.562      0.575   -1.52e+06    2.73e+06
PPG           6.301e+04      2e+04      3.150      0.002    2.33e+04    1.03e+05
Avg          -5.174e+04   2.28e+04     -2.272      0.025   -9.69e+04   -6577.638
Win_Perc     -1.675e+06   9.09e+05     -1.843      0.068   -3.48e+06    1.28e+05
GSR           7500.1741   7949.921      0.943      0.348   -8264.818    2.33e+04
Capacity        31.2074      5.880      5.308      0.000      19.548      42.867
ACC           8.929e+05   2.71e+05      3.293      0.001    3.55e+05    1.43e+06
AAC          -5.475e+05   2.95e+05     -1.855      0.066   -1.13e+06    3.79e+04
Big12         1.315e+06   3.12e+05      4.209      0.000    6.95e+05    1.93e+06
Big10         1.237e+06   2.93e+05      4.218      0.000    6.56e+05    1.82e+06
CUSA         -9.186e+05   2.54e+05     -3.616      0.000   -1.42e+06   -4.15e+05
Independents -7.735e+05   4.52e+05     -1.712      0.090   -1.67e+06    1.22e+05
MidAmerican  -7.368e+05    2.8e+05     -2.636      0.010   -1.29e+06   -1.83e+05
MountainWest -8.228e+05   2.83e+05     -2.910      0.004   -1.38e+06   -2.62e+05
Pac12         4.054e+05    2.8e+05      1.448      0.151    -1.5e+05    9.61e+05
SEC           1.316e+06   3.03e+05      4.338      0.000    7.14e+05    1.92e+06
SunBelt      -7.648e+05   3.43e+05     -2.227      0.028   -1.45e+06   -8.39e+04
==============================================================================
Omnibus:                        1.237   Durbin-Watson:                   1.851
Prob(Omnibus):                  0.539   Jarque-Bera (JB):                0.784
Skew:                           0.014   Prob(JB):                        0.676
Kurtosis:                       3.395   Cond. No.                     4.96e+20
==============================================================================
```

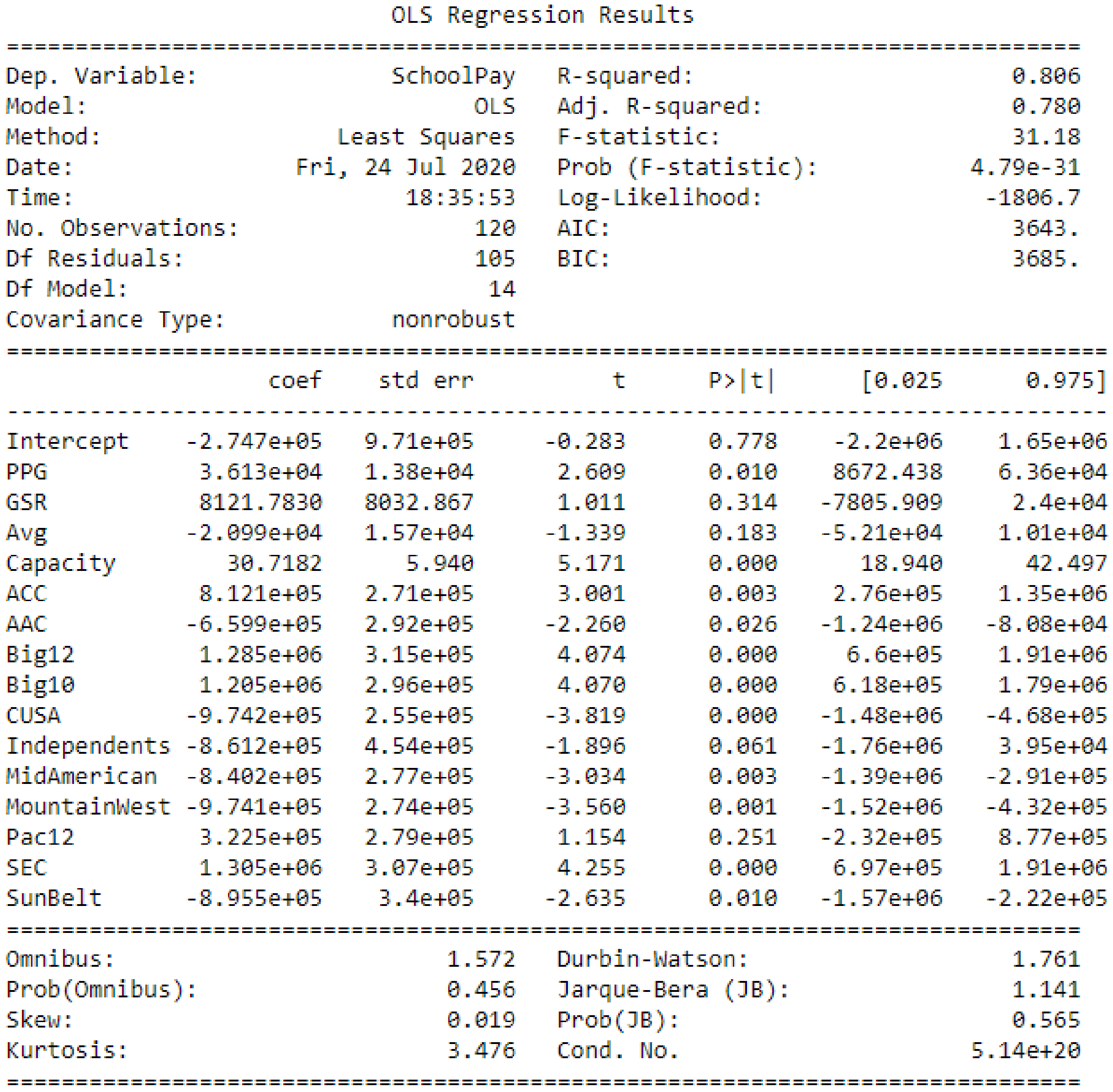
Model 1 is a linear regression model derived from the statsmodels package in Python. Already, there is an obvious problem in this model. Namely, the Win Percentage coefficient is negative. This should not be the case as it was seen in the previous analysis that salary should increase as win percentage increases. To test for potential multicollinearity, a correlation matrix was created to examine the relationship between the independent variables.

Jose Reyes
7.23.2020
IST 718



There appears to be a high correlation between Average points allowed and Win Percentage likely because they are measuring similar performances. For the next model iteration, the Win Percentage variable will be dropped as it has a higher p value between the two.

Jose Reyes
7.23.2020
IST 718



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SchoolPay   R-squared:                       0.806
Model:                            OLS   Adj. R-squared:                  0.780
Method:                 Least Squares   F-statistic:                     31.18
Date:                Fri, 24 Jul 2020   Prob (F-statistic):           4.79e-31
Time:                        18:35:53   Log-Likelihood:                -1806.7
No. Observations:                 120   AIC:                             3643.
Df Residuals:                     105   BIC:                             3685.
Df Model:                          14
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept    -2.747e+05   9.71e+05     -0.283      0.778    -2.2e+06    1.65e+06
PPG           3.613e+04   1.38e+04      2.609      0.010    8672.438    6.36e+04
GSR           8121.7830   8032.867      1.011      0.314   -7805.909     2.4e+04
Avg          -2.099e+04   1.57e+04     -1.339      0.183   -5.21e+04    1.01e+04
Capacity        30.7182      5.940      5.171      0.000      18.940      42.497
ACC           8.121e+05   2.71e+05      3.001      0.003    2.76e+05    1.35e+06
AAC          -6.599e+05   2.92e+05     -2.260      0.026   -1.24e+06   -8.08e+04
Big12         1.285e+06   3.15e+05      4.074      0.000     6.6e+05    1.91e+06
Big10         1.205e+06   2.96e+05      4.070      0.000    6.18e+05    1.79e+06
CUSA         -9.742e+05   2.55e+05     -3.819      0.000   -1.48e+06   -4.68e+05
Independents -8.612e+05   4.54e+05     -1.896      0.061   -1.76e+06    3.95e+04
MidAmerican  -8.402e+05   2.77e+05     -3.034      0.003   -1.39e+06   -2.91e+05
MountainWest -9.741e+05   2.74e+05     -3.560      0.001   -1.52e+06   -4.32e+05
Pac12         3.225e+05   2.79e+05      1.154      0.251   -2.32e+05    8.77e+05
SEC           1.305e+06   3.07e+05      4.255      0.000    6.97e+05    1.91e+06
SunBelt      -8.955e+05    3.4e+05     -2.635      0.010   -1.57e+06   -2.22e+05
==============================================================================
Omnibus:                        1.572   Durbin-Watson:                   1.761
Prob(Omnibus):                  0.456   Jarque-Bera (JB):                1.141
Skew:                           0.019   Prob(JB):                        0.565
Kurtosis:                       3.476   Cond. No.                     5.14e+20
==============================================================================
```

The R-squared value in this iteration dropped slightly, but we can now intuitively explain the coefficient values. As PPG increases and Avg decreases, salary increases. Capacity has a relatively small coefficient which makes sense considering this value ranges widely by thousands between schools. The coefficient magnitudes for the conferences also match what was seen in the initial exploratory analysis. Schools in the SEC, Big12, and Big10 provide the highest increases to coach salary while CUSA, Mountain West, and Sun Belt provide the lowest.

Jose Reyes
7.23.2020
IST 718

Model 2 results:

- Coefficients:

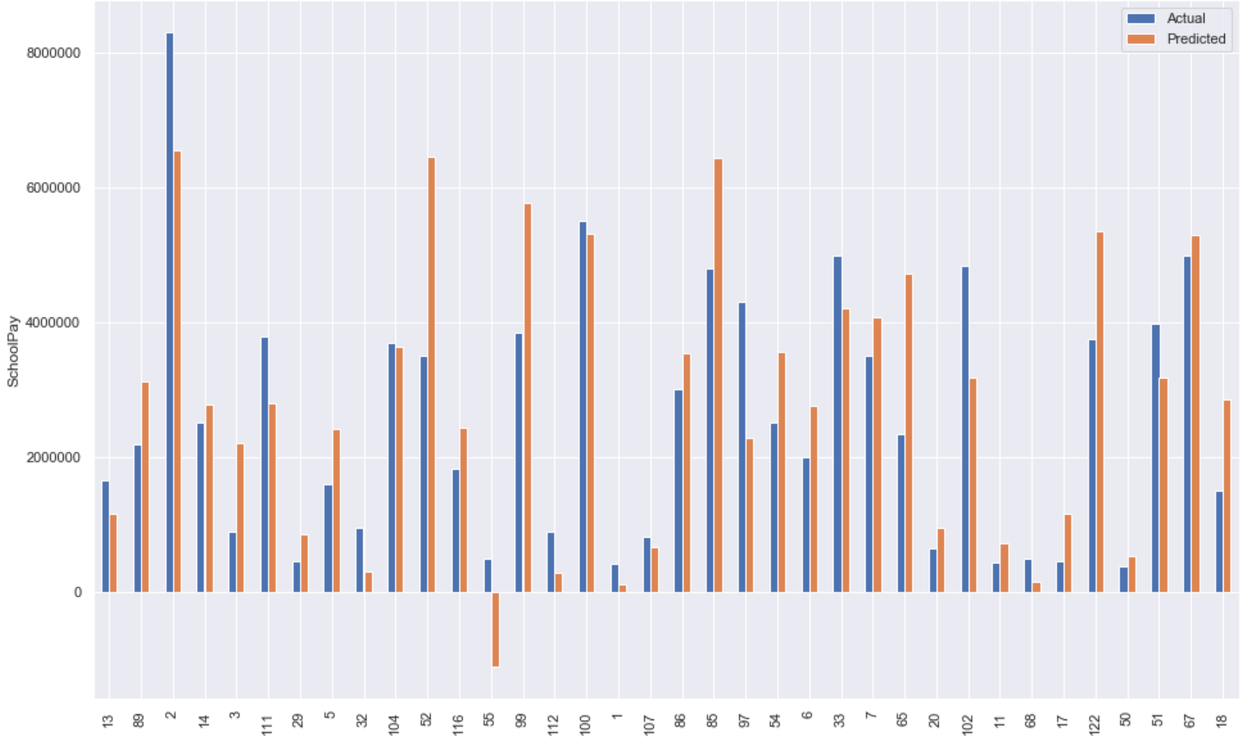
  PPG: 3.02941838e+04
  GSR: 2.43299773e+03
  Avg: -2.94258090e+04
  Capacity: 3.76225100e+01
  ACC: 7.67176295e+05
  AAC: -5.69623866e+05
  Big12: 9.73163955e+05
  Big10: 1.44033574e+06
  CUSA: -8.61329999e+05
  Independents: -1.11667077e+06
  Mid-American: -5.72914929e+05
  MountainWest: -1.03386324e+06
  PAC 12: 2.25393007e+05
  SEC: 1.49528411e+06
  SunBelt: -7.46950302e+05

- Model 2 R-squared: 0.855

A second linear regression model was created this time from the Scikit-learn package using the same inputs as Model 1. In this case, the data set was split 70/30% for model training and testing respectively. The test results produced an R-squared value of 0.855, which is relatively high and indicates that this is a pretty strong model. In other words, 85% of the variance can be explained by this linear regression model. The coefficients also very closely match the results shown in Model 1.

Jose Reyes
7.23.2020
IST 718

The following graph shows the model predictions compared to the actual salary values. Despite the relatively high R-squared value, it should be noted that this model's imperfection is clearly displayed for one coach where the model predicted a negative salary.



Because model 2 produced a higher R-squared value and thus has a higher goodness-of-fit, it will be used to make a salary prediction for the next Syracuse coach. Using the inputs in the model, the Syracuse coach's salary should be $2,934,224 compared to the current salary of $2,401,206. This comes as a surprise considering Syracuse has generally lower than average values for most of the independent variables. It is evident that the model places a lot of significance on the school's conference as this appears to have outweighed the other variables.

If Syracuse moves to the Big10, the model predicts that the Syracuse coach's salary should be $3,608,383. This makes sense as schools in the Big10 generally have higher salaries than the ACC. If Syracuse moves to the Big East, which is a non-football conference, it is likely that they would play as an Independent program. This would allow Syracuse to continue playing in the Football Bowl Subdivision (FBS) and stay eligible for a Bowl Championship. Using Model 2 to predict the next Syracuse coach's salary as an independent school, the salary would be $1,050,377. Again, the lower value makes sense as independent schools generally have lower salaries compared to ACC and all other P5 schools.

Clearly, the model produces rather large upswings and downswings in salary when a school switches conference. It is because of this that conference appears to be the single most significant variable in the model. This was observed when the model predicted a higher salary for the Syracuse coach than the current salary despite having below average performances in nearly all other metrics. The model also predicts large salary differences if a school switches from a P5 conference to a non-P5 conference. Even within P5 conferences, the model predicted an increase of around $800,000 if Syracuse moves to the Big10. This is additionally displayed by the large magnitudes of the dummy variable coefficients seen in the regression results.

On the other hand, GSR appears to have the least significant impact on salary size based on the model coefficients. This would make sense when determining a coach's salary as it is difficult to attribute this metric to a coach's performance. Similarly, the exploratory data analysis showed a less pronounced relationship between GSR and salary size compared to other variables.

Conclusion

The next Syracuse coach's salary was predicted using exploratory data analysis and linear modelling. This was first done by creating a single data frame with information on each Division 1 Football programs including their records, stadium capacity, and graduation rates. Each variable was explored and evaluated to identify if a relationship exists with coach salaries.

The exploratory analysis showed a positive relationship between the coach's salary and several indicators of a strong football program including a higher win percentage, points per game, stadium size, and lower points allowed per game. The school's athletic conference also has a relationship with salary as coaches in the P5 conferences have higher salaries than others.

From the modelling, the regression results largely matched the findings in the exploratory analysis. Overall, the school's conference appears to be the most significant contributor to a coach's salary size while GSR is the least. Of the 2 models, Model 2 provided the higher R-squared value and was used to predict the next Syracuse's coach salary. In conclusion, based on all the analysis and modelling provided, the next Syracuse coach's salary should be $2,934,224.