Mario Ruiz

Professor Fox

IST 718

Lab 1 – Coaching Salary

Introduction

Coaching salaries are often topics of interest within many educational institutions. The privilege of education is generally an attempt to foster intellectual growth, to better prepare students for the eventual workplace. Thus, one may wonder whether the purpose of athletics at these educational institutions is to build character, and interpersonal problem-solving skill. Then, should a coaches' salary should be contingent on graduation success rate, along with season performance? However, in some cases, if the ability of a team to generate income is the major factor, should stadium attendance, and team performance be major factors on the coaches' salary? Before attempting to tackle such questions, understanding the driving mechanisms is a fundamental first step.

In this study both linear regression, and ordinary least squares will be used to predict a recommended salary for the Syracuse football team. Therefore, several data sources will need to be aggregated to allow a normalized comparison between various NCAA Division I coaches. Then, out of the chosen factors, a determination will be made, indicating which were factors were significant for the given model.

Analysis

Data Preparation:

Four csv datasets were used, three were manually created using various sources:

- ❖ coaches: supplied list of division 1 football coaches
- ❖ season_2017[1]
- ❖ ncaa football stadiums[2]
- ❖ graduation rates[3]

Typical conversion techniques were implemented, including conversion to lowercased, replacing non-numeric characters to empty spaces, along with coercing numeric values

[1] Google search: ncaa 2017 football
[2] https://github.com/gboeing/data-visualization/blob/master/ncaa-football-stadiums/data/stadiums-geocoded.csv
[3] http://www.ncaa.org/about/resources/research/graduation-rates

from string. However, since multiple datasets were loaded to their own dataframes, each eventually needed to be joined. Therefore, standardization required column names to be consistently coded:

```
stadium['school'] = stadium['school'].replace(['ucf'], 'central florida')
stadium['school'] = stadium['school'].replace(['usf'], 'south florida')
stadium['school'] = stadium['school'].replace(['utsa'], 'texas-san antonio')
stadium['school'] = stadium['school'].replace(['byu'], 'brigham young')
stadium['school'] = stadium['school'].replace(['utep'], 'texas-el paso')
stadium['school'] = stadium['school'].replace(['tcu'], 'texas christian')
stadium['school'] = stadium['school'].replace(['unlv'], 'nevada-las vegas')
stadium['school'] = stadium['school'].replace(['smu'], 'southern methodist')
stadium['school'] = stadium['school'].replace(['niu'], 'northern illinois')
stadium['school'] = stadium['school'].replace(['miami (oh)'], 'miami (ohio)')
stadium['school'] = stadium['school'].replace(['fiu'], 'florida international')
stadium['school'] = stadium['school'].replace(['umass'], 'massachusetts')
stadium['school'] = stadium['school'].replace(['yale bulldogs'], 'connecticut')
```

Techniques such as these, allowed the `stadium` dataframe to join with the `coaches` dataframe, then with the `grad_rate`, and `season_2017` dataframes using a common `school` column. Any rows not having a common column, would not be included in the overall joined dataframe. Once all three dataframe were joined together, a single `merged_df` was used for successive analysis.

However, since requirements for this study involved the 2006 student athlete cohort, stadiums expanded after 2006 were removed. Otherwise, the parameters for the graduation rates would not be relevant.

Finally, the train was created using 2/3 of the original merged_df dataset, while the remaining 1/3 was reserved for testing. This allowed the sklearn `LinearRegression`, as well as the scipy `ols` (ordinary least squares) to be implemented for model fitting. Originally, several independent variables were used for training:

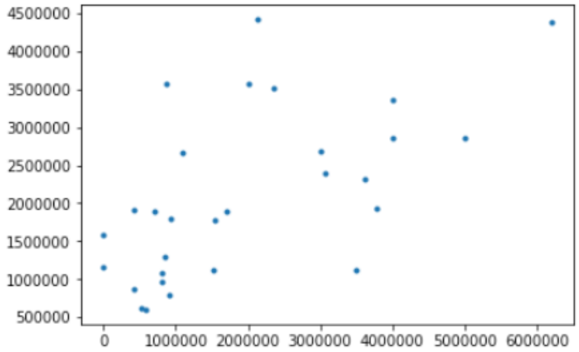
- capacity: football stadium capacity
- gsr: graduation success rate
- fgr: federal graduation rate
- win: total 2017 season wins for a given team
- loss: total 2017 season losses for a given team
- pct: ratio of win / loss

However, the independent variables were reduced to the following:

- capacity: football stadium capacity
- gsr: graduation success rate
- pct: ratio of win / loss

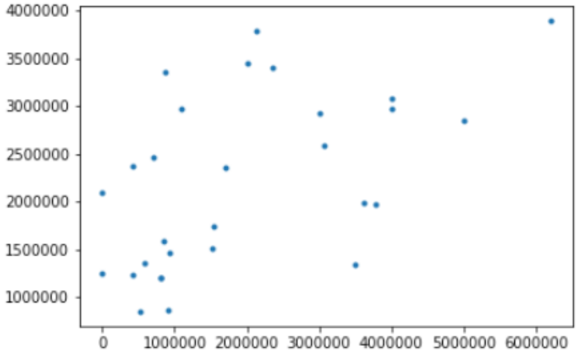
Results

The determined `LinearRegression` fit generated an r-squared of `0.330`:

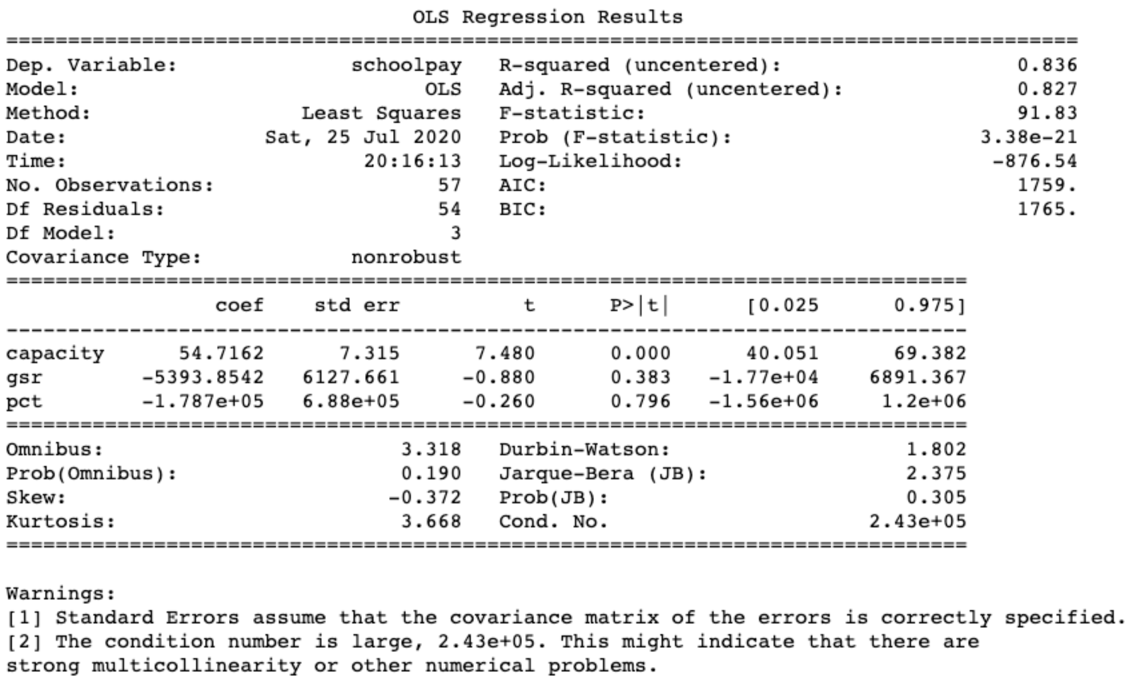


This indicates the model is not good at accounting variance of coaches' salary with the selected independent variables. Using the associated model, the predicted salary for a Syracuse football coach is estimated at $`2,081,669.53`.

Next, the `ols` model was computed using the same factors:



This method generated better insight, by providing measures indicating which components of the model were significant:



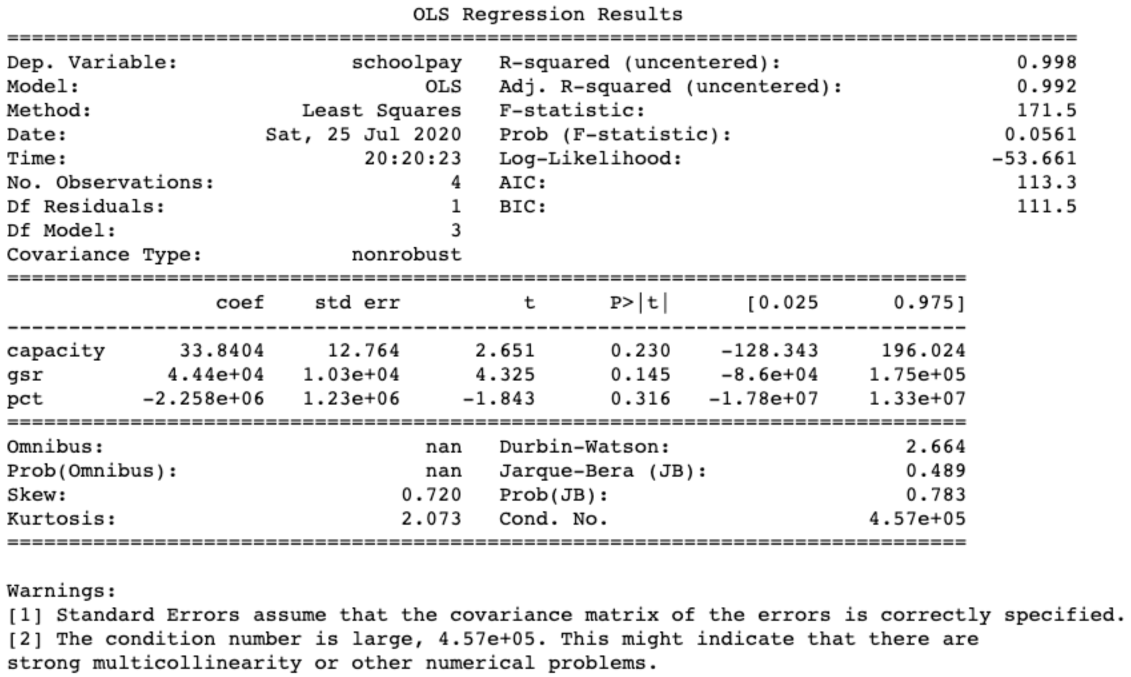
```
                            OLS Regression Results
=======================================================================================
Dep. Variable:              schoolpay   R-squared (uncentered):                   0.836
Model:                            OLS   Adj. R-squared (uncentered):              0.827
Method:                 Least Squares   F-statistic:                              91.83
Date:                Sat, 25 Jul 2020   Prob (F-statistic):                    3.38e-21
Time:                        20:16:13   Log-Likelihood:                         -876.54
No. Observations:                  57   AIC:                                      1759.
Df Residuals:                      54   BIC:                                      1765.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
capacity      54.7162      7.315      7.480      0.000      40.051      69.382
gsr        -5393.8542   6127.661     -0.880      0.383   -1.77e+04    6891.367
pct        -1.787e+05   6.88e+05     -0.260      0.796   -1.56e+06     1.2e+06
==============================================================================
Omnibus:                        3.318   Durbin-Watson:                   1.802
Prob(Omnibus):                  0.190   Jarque-Bera (JB):                2.375
Skew:                          -0.372   Prob(JB):                        0.305
Kurtosis:                       3.668   Cond. No.                     2.43e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.43e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Specifically, the stadium `capacity`, along with `gsr` for a given team are significant. The `pct` wins was not significant for the overall model, while the above `3.38e-21` Prob (F-Statistic) indicates a significant regression model. Most importantly, the `ols` model generated an r-squared of `0.827`. This indicates a significantly better result than the `LinearRegression`. Furthermore, the `1.802` for the Durbin-Watson indicates limited autocorrelation exists between the selected factors. Therefore, the selected factors to train were not redundant. The computed prediction of what a Syracuse football coach was found to be $`2,219,865.00`

Since the `ols` method proved to generate better results, this implementation was used for the hypothetical scenario if Syracuse was in the Big 10. The dataset used was significantly reduced, by filtering on the Big 10 conference:

```
train_big10, test_big_10 = train_test_split(merged_df[merged_df['conference']
== 'big ten'], test_size=0.33)
```

The corresponding model generated a nonsignificant probability F-Statistic, with an overall salary prediction of $3,437,561. Additionally, the corresponding factors were all nonsignificant:



```
                            OLS Regression Results
=======================================================================================
Dep. Variable:             schoolpay   R-squared (uncentered):                   0.998
Model:                           OLS   Adj. R-squared (uncentered):              0.992
Method:                Least Squares   F-statistic:                              171.5
Date:               Sat, 25 Jul 2020   Prob (F-statistic):                      0.0561
Time:                       20:20:23   Log-Likelihood:                         -53.661
No. Observations:                  4   AIC:                                      113.3
Df Residuals:                      1   BIC:                                      111.5
Df Model:                          3
Covariance Type:           nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
capacity      33.8404     12.764      2.651      0.230    -128.343     196.024
gsr          4.44e+04   1.03e+04      4.325      0.145    -8.6e+04    1.75e+05
pct        -2.258e+06   1.23e+06     -1.843      0.316   -1.78e+07    1.33e+07
==============================================================================
Omnibus:                         nan   Durbin-Watson:                   2.664
Prob(Omnibus):                   nan   Jarque-Bera (JB):                0.489
Skew:                          0.720   Prob(JB):                        0.783
Kurtosis:                      2.073   Cond. No.                     4.57e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.57e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Instead of filtering the dataset on the Big 10 conference, it may have been more appropriate to factor the column into integer values, then retain the column during train. The latter implementation had too few coaches to accurately model the given scenario. Furthermore, since the original coach dataset does not contain any Big East coaches, a similar hypothetical question of a Syracuse coach being in the Big East cannot be estimated.

To better improve the overall modeling, rows with missing coaching salaries (normalized to $0), could have been forced to the test set. Rather in this study, the test and train set were randomly distributed.

Conclusions

Choosing an appropriate regression model is often an important task when attempting to make a prediction. As indicated in this study, the linear regression model using sklearn `LinearRegression` generates an r-squared significantly worse than the ordinary least squares available through the `statsmodel` package. Both techniques though having different results, internally implement the ordinary least squares model. Therefore, given more time, it would be interesting to determine which parameters could be adjusted using the `sklearn`, and whether model performance could be improved. Additionally, both techniques suggest that a Syracuse football coach should make roughly $2M.

When trying to understand the discrepancy between the two salaries, its difficult to argue that the team record was not accounted for, since in 2017, the football team had a 4-8 record. Additionally, the graduation success rate as 77 for 2006 was about average relative to other teams used for the analysis. However, when reviewing the overall model holistically, it does not seem like a reasonable approach. Specifically, using the 2006 graduate success rate, while using the 2017 (last years) season record seem disjoint. Having this level of difference, would have a large impact on the overall model. Instead, having the two factors both represent 2006, or 2017 would be more appropriate when generating the corresponding model.

Lastly, more data for the coaches’ dataset would have improved the overall model. Specifically, no coaching information was provided for the Big East conference in the original coaches’ dataset. Therefore, attempting to answer any related question would not have been possible without additional data aggregation.