Jagannathan Govindan

IST 718

Lab #1

**Introduction:**

A college football program can be turned around by the hiring of good or great coaches. Additional factors do play in the winning percentages like recruiting, conferences and also schedule. However in general coaches play a great role in NCAA. So determining an optimum salary for a coach becomes an interesting data analytical problem.

In this lab experiment both linear regression and OLS will be used to predict a recommended salary for the Syracuse football team. Data from different sources are combined, cleaned and used in this and their sources are credited in the end.

**Obtain and Scrub:**

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

* coaches: supplied list of division 1 football coaches
* season\_2017[[1]](#footnote-1)
* NCAA football stadiums[[2]](#footnote-2)
* graduation rates[[3]](#footnote-3)

In this phase, usually obtain and scrub techniques were followed like taking care of NA characters, converting from string to numeric and also lower case and upper cases conversions. Since multiple datasets are involved, standardized column names are used as below:

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')

The above process helped in joining the dataframes **coaches**(provided as part of the Assignment), then **grad\_rate** and **season\_2017** by school name. once done, the merged dataframe **merged\_df** was used for running regression and OLS.

As this assignment pertains to GSR and FSR for 2006 cohorts, we decided to ignore the stadiums expanded or renovated after 2006 which will make the analysis relevant.

Finally, the train was created using 2/3 of the original **merged\_df** dataset, while the remaining 1/3 was reserved for testing. Based on this train and test dataset both Linear and OLS models were ran and looked for best fit based on the R square value. Originally, several independent variables were used for training:

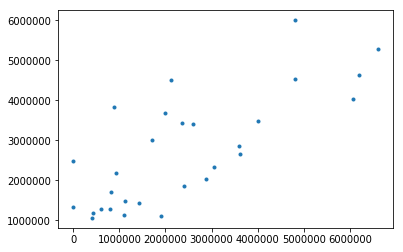
* capacity: football stadium capacity
* gsr: graduation success rate
* fgr: federal graduation rate
* win: total 2017 season wins for a given team
* loss: total 2017 season losses for a given team
* pct: ratio of win / loss

However, the independent variables were reduced to the following:

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

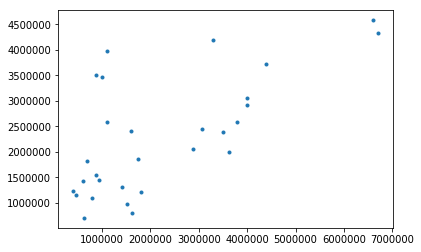
**Results:**

The determined **LinearRegression** fit generated an R-squared of 0.512:



From the R-squared value, this model seems to be not a great fit for the variables involved. But the predicted salary for a Syracuse football coach is estimated at $1,863,383.60.

Next, the **OLS** model was computed using the same factors:



This model had a better R-squared value as shown below:

OLS Regression Results

==================================================w============================

Dep. Variable: schoolpay R-squared: 0.814

Model: OLS Adj. R-squared: 0.804

Method: Least Squares F-statistic: 78.90

Date: Sun, 28 Oct 2018 Prob (F-statistic): 9.84e-20

Time: 01:00:55 Log-Likelihood: -876.87

No. Observations: 57 AIC: 1760.

Df Residuals: 54 BIC: 1766.

Df Model: 3

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

capacity 53.6009 7.966 6.729 0.000 37.630 69.572

gsr -1.436e+04 6615.861 -2.171 0.034 -2.76e+04 -1098.499

pct 9.094e+05 6.66e+05 1.365 0.178 -4.26e+05 2.24e+06

==============================================================================

Omnibus: 0.365 Durbin-Watson: 2.424

Prob(Omnibus): 0.833 Jarque-Bera (JB): 0.050

Skew: 0.048 Prob(JB): 0.976

Kurtosis: 3.109 Cond. No. 2.26e+05

==============================================================================

Other inferences from this model is the stadium capacity plays a major role than win % or GSR which is an interesting finding and the predicted salary for Syracuse football coach was $1,837,076.

This model had the better R-squared value for the variables, so it seems to be better than the Linear regression. It was used for further progression including the assumption that Syracuse in Big10. Again it is done by splitting the data as below:

train\_big10, test\_big\_10 = train\_test\_split(merged\_df[merged\_df['conference'] == 'big ten'], test\_size=0.33)

The model results are below and the predicted salary for Syracuse football program coach in case of Big10 is $3,437,561.

OLS Regression Results

==============================================================================

Dep. Variable: schoolpay R-squared: 0.992

Model: OLS Adj. R-squared: 0.969

Method: Least Squares F-statistic: 42.15

Date: Sun, 28 Oct 2018 Prob (F-statistic): 0.113

Time: 01:41:52 Log-Likelihood: -56.963

No. Observations: 4 AIC: 119.9

Df Residuals: 1 BIC: 118.1

Df Model: 3

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

capacity 21.0401 21.313 0.987 0.504 -249.763 291.843

gsr 2.861e+04 5.64e+04 0.508 0.701 -6.88e+05 7.45e+05

pct 5.94e+05 6.84e+06 0.087 0.945 -8.63e+07 8.75e+07

==============================================================================

Omnibus: nan Durbin-Watson: 1.994

Prob(Omnibus): nan Jarque-Bera (JB): 0.607

Skew: -0.004 Prob(JB): 0.738

Kurtosis: 1.092 Cond. No. 1.38e+06

==============================================================================

**Conclusions:**

From the two models ran for this assignment, OLS seems to be a better fit assuming only the few independent variables taken. This same model can also be tried with other varialbles and also we can try to fit this data to other models like Random Forest but due to time constraint I couldn’t try them. Syracuse program is in ACC and

**Credits:**

NCAA 2017 football

2 <https://github.com/gboeing/data-visualization/blob/master/ncaa-football-stadiums/data/stadiums-geocoded.csv> -- This one was referred by my team mate

3 <http://www.ncaa.org/about/resources/research/graduation-rates> -- Division 1

**Python code:**



1. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)
3. [↑](#footnote-ref-3)