NCAA Football Coaches Salary Analysis

Given a set of data on NCAA Football coaches salaries, we were asked to answer the following questions:

- What is the recommended salary for the Syracuse football coach?
- What would his salary be if we were still in the Big East? What if we went to the Big Ten?
- What schools did we drop from our data, and why?
- What effect does graduation rate have on the projected salary?
- How good is our model?
- What is the single biggest impact on salary size?

To answer these questions, we took the given set of data, and merged it with:

1. Stadium capacity data, pulled from Wikipedia:
    a. https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_stadiums
2. Graduation Rate data, pulled from NCAA statistics:
    a. 2018RES_File5-DISquadAggregationSA.txt
3. The 2018 regular season W/L records, pulled from the NCAA website:
    a. https://www.cbssports.com/college-football/standings

This Team/School and Conference names in each of these data sets did not all perfectly match. To resolve this, we used fuzzy matching from the **fuzzywuzzy** Python package, which allowed us to match names in the datasets that were similar, using a score to judge the similarity. This fuzzy matching process allowed us to build a bridge table, mapping the Team/Conference names of each additional data set back to the School/Conference names in the original salary data. Any entries in the bridge tables that were found to be incorrect were manually corrected before mapping the dependent data values back to the salary data.

Once the data was clean and compiled into a single dataset, the data for Baylor, Brigham Young, Rice, and Southern Methodist were dropped due to not having valid Salary data. Also, the Air Force, Army, Charlotte and New Mexico State data were dropped due to not having valid Graduation Rate data.

A linear-regression model was to predict the salary paid by the school to the coach (SchoolPay) was built using the W/L Percentage from 2018, the Stadium capacity, and the Conference as predictors. This is a fairly good model with an Adjusted R-Squared of 0.78, meaning that 78% of the variation in SchoolPay is explained by the selected predictor variables.

[35]:

OLS Regression Results

Dep. Variable:	SchoolPay	**R-squared:**	0.806
Model:	OLS	**Adj. R-squared:**	0.784
Method:	Least Squares	**F-statistic:**	37.28
Date:	Sat, 20 Jul 2019	**Prob (F-statistic):**	6.86e-33
Time:	23:45:17	**Log-Likelihood:**	-1820.6
No. Observations:	121	**AIC:**	3667.
Df Residuals:	108	**BIC:**	3703.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-5.436e+05	3.84e+05	-1.416	0.160	-1.3e+06	2.17e+05
Conference[T.ACC]	1.177e+06	3.71e+05	3.175	0.002	4.42e+05	1.91e+06
Conference[T.Big 12]	1.57e+06	4.16e+05	3.770	0.000	7.45e+05	2.4e+06
Conference[T.Big Ten]	1.63e+06	3.92e+05	4.160	0.000	8.53e+05	2.41e+06
Conference[T.C-USA]	-6.046e+05	3.77e+05	-1.604	0.112	-1.35e+06	1.43e+05
Conference[T.Ind.]	-4.855e+05	5.79e+05	-0.838	0.404	-1.63e+06	6.63e+05
Conference[T.MAC]	-4.858e+05	3.86e+05	-1.257	0.211	-1.25e+06	2.8e+05
Conference[T.Mt. West]	-5.175e+05	3.84e+05	-1.347	0.181	-1.28e+06	2.44e+05
Conference[T.Pac-12]	6.545e+05	3.83e+05	1.710	0.090	-1.04e+05	1.41e+06
Conference[T.SEC]	1.544e+06	4.12e+05	3.749	0.000	7.28e+05	2.36e+06
Conference[T.Sun Belt]	-5.886e+05	3.99e+05	-1.475	0.143	-1.38e+06	2.03e+05
Pct	1.483e+06	3.91e+05	3.797	0.000	7.09e+05	2.26e+06
StadiumCapacity	33.9345	5.415	6.267	0.000	23.202	44.667

Omnibus:	1.785	**Durbin-Watson:**	1.953
Prob(Omnibus):	0.410	**Jarque-Bera (JB):**	1.277
Skew:	0.175	**Prob(JB):**	0.528
Kurtosis:	3.361	**Cond. No.**	6.86e+05

Figure 1: Final Selected Model

The single biggest impact to Salary seems to be Conference. The baseline Conference was assumed to be the American Athletic Conference (AAC), and just by switching to another conference like the Big Ten the Salary increases about $1.7 million dollars on average. Conversely, switching to Conference USA (C-USA) can decrease the Salary by about $600,000 dollars on average.

This model estimates that the Syracuse football coach should receive approximately $3.4 million dollars in annual salary, versus the current salary of approximately $2.4 million dollars.

If Syracuse switched to the Big Ten, then this model would predict a Salary of approximately $3.9 million dollars. If Syracuse were to switch back to the Big East Conference, then they would not compete in Division I football, but I believe it is fair to classify those teams similarly to the Independent teams in the available data. This model would then predict the Salary to be about $1.8 million dollars.

A model was built including the Graduation Rate data, but those variables were not found to be statistically significant, and were removed from the final selected model. This was because the p-values were high, and also the high and low values for the predicted coefficient where on either side of zero, indicating that the impact of the graduation rate variables could not be shown to be non-zero with a high degree of probability.

[32]:
```
# specify a simple model with bobblehead entered last
my_model = str('SchoolPay ~ FSR + GSR + Pct + StadiumCapacity')

# fit the model to the training set
model_fit = smf.ols(my_model, data=model_data).fit()
model_fit.summary()
```

[32]:

OLS Regression Results

Dep. Variable:	SchoolPay	**R-squared:**	0.688
Model:	OLS	**Adj. R-squared:**	0.677
Method:	Least Squares	**F-statistic:**	63.96
Date:	Sat, 20 Jul 2019	**Prob (F-statistic):**	1.86e-28
Time:	23:45:17	**Log-Likelihood:**	-1849.2
No. Observations:	121	**AIC:**	3708.
Df Residuals:	116	**BIC:**	3722.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-2.04e+06	8.11e+05	-2.515	0.013	-3.65e+06	-4.33e+05
FSR	6244.5453	1.37e+04	0.456	0.650	-2.09e+04	3.34e+04
GSR	2712.9562	1.62e+04	0.168	0.867	-2.93e+04	3.47e+04
Pct	1.179e+06	4.87e+05	2.419	0.017	2.14e+05	2.14e+06
StadiumCapacity	63.4585	4.404	14.410	0.000	54.736	72.181

Omnibus:	4.107	**Durbin-Watson:**	2.055
Prob(Omnibus):	0.128	**Jarque-Bera (JB):**	3.883
Skew:	-0.265	**Prob(JB):**	0.143
Kurtosis:	3.699	**Cond. No.**	4.73e+05

Figure 2: Model With All Numeric Predictors