Jiebin (Alex) Zhu

IST 718 – Lab 1 College Football Coach Salary Recommendation

This document is created as a supplement to the code, mostly to explain about the visualization and models. It will also contain some recommendations and answers to the lab questions. Most of the data cleaning will be explained within the code that is attached with this submission.

Constraints

This lab contained a few constraints that should be noted as it could've affected the results, despite already taking appropriate measures to mitigate the risks here:

- Inconsistent naming format across different data sets which caused some friction with merging the datasets together, resulting in additional observations being removed and potentially introducing some risk of accidentally merging incorrectly that's inevitable.
- Missing data in several columns causing some of the observations to be removed.
- Published date on gather data sets could be different causing some data inconsistency such as school conferences.

Data Exploration

From the initial glance of the data **(Figure 1.1)**, it seems that there are outliers within the conferences. However, the conferences that had a significantly higher coach salary seem to belong to the Power Five conferences **(SEC, Pac-12, ACC, Big Ten, Big 12).** Conferences could be potentially an important variable for predicting the coaches' salary.

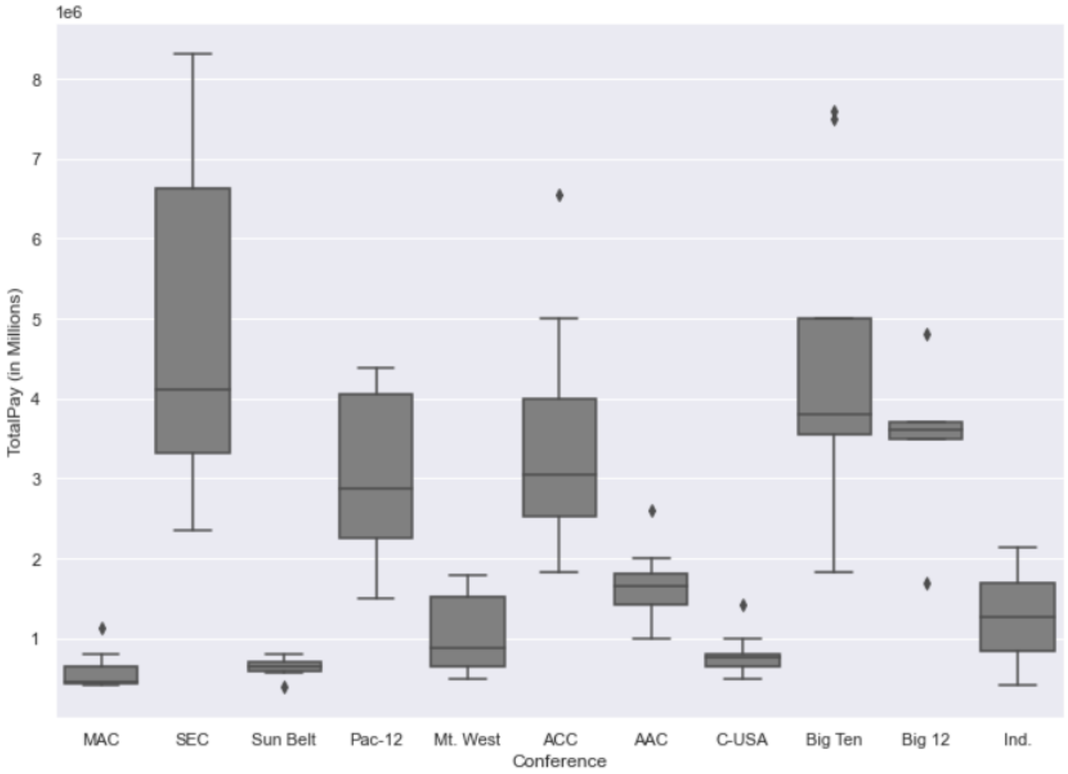


Figure 1.1

With that said, the only next reasonable thing to do here is to redo the visualization to identify outliers with that new variable. **(Figure 1.2)** Only a single outlier was identified and removed here after redoing the visualization. This also means that power 5 conference variable would potentially play a huge role in determining the coaches' salary. Since the power 5 conferences contain elite schools, it would make sense that their coaches are also well compensated for that.

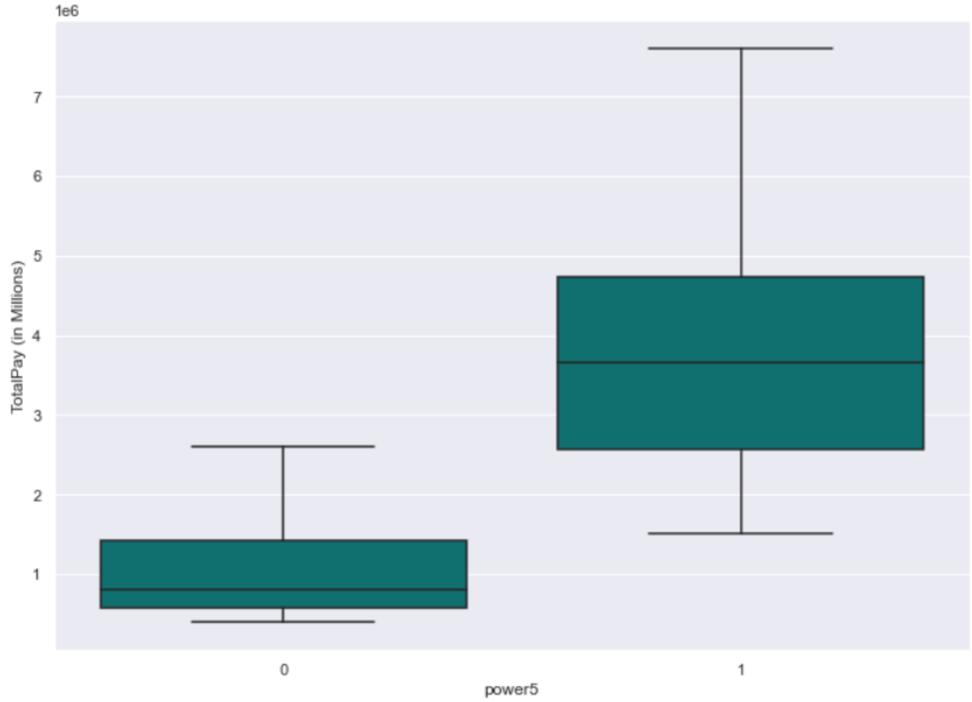


Figure 1.2

In a scatter plot **(Figure 1.3)** that was created to identify the relationship between the stadium seating capacity and the coaches' salary it is very noticeable that there's some sort of correlation. Schools that tend to have stadiums that are greater, also have higher paying salaries for their coaches. Another interesting finding here is that those schools with larger stadiums also seem to be the schools in the power 5 conferences. There are many ways to interpret this:

- School has a lot of money to pay their coaches and win more so they can expand their stadium because they could see an influx of fans.
- Stadium is large enough to seat more fans and increase the school's revenue. This allows them to pay their coach more.
- The large stadium can seat a lot of fans and teams win more because of the cheering, which allows the school to be placed in power 5 causing the coaches' pay to increase.
- Etc.

However, there isn’t a correct answer here and it is open for interpretation.

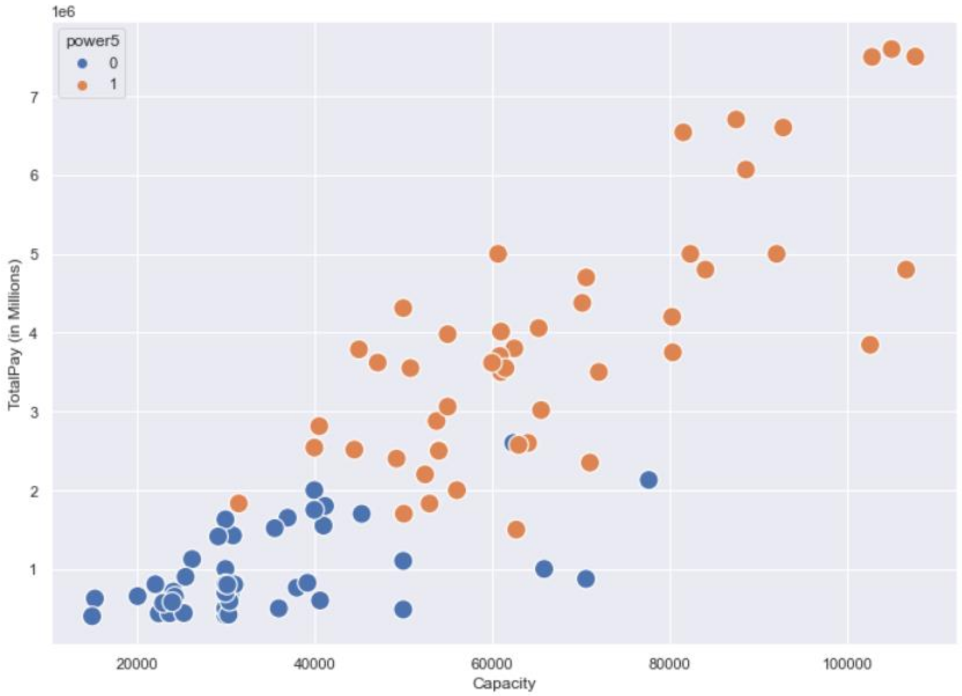


Figure 1.3

Since the one of the interpretations is that the schools do really well and are known as the power 5 conferences would have higher coach pay, next looking at the most recent win percentages would be the best approach **(Figure 1.4)**. However, it seems that there isn’t really a correlation with the most games and we might not even need this variable later on.

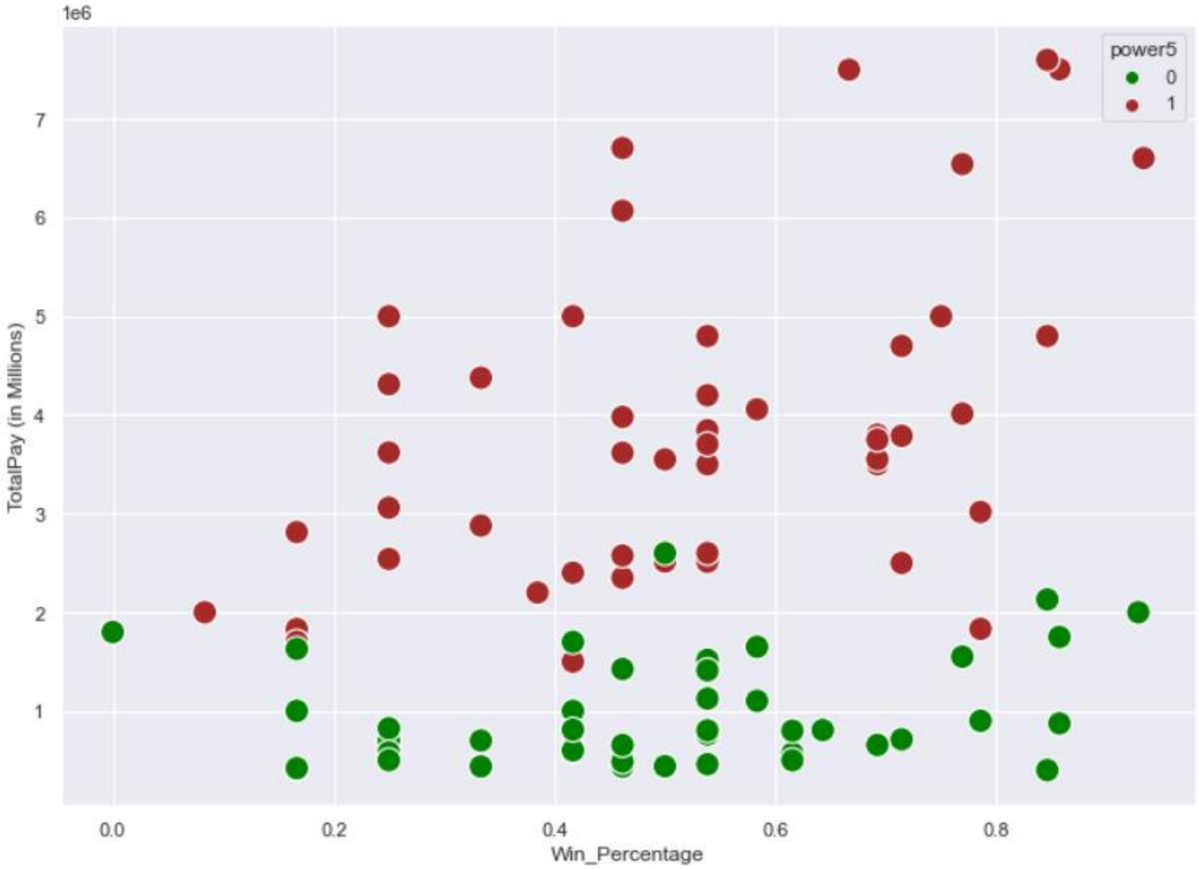
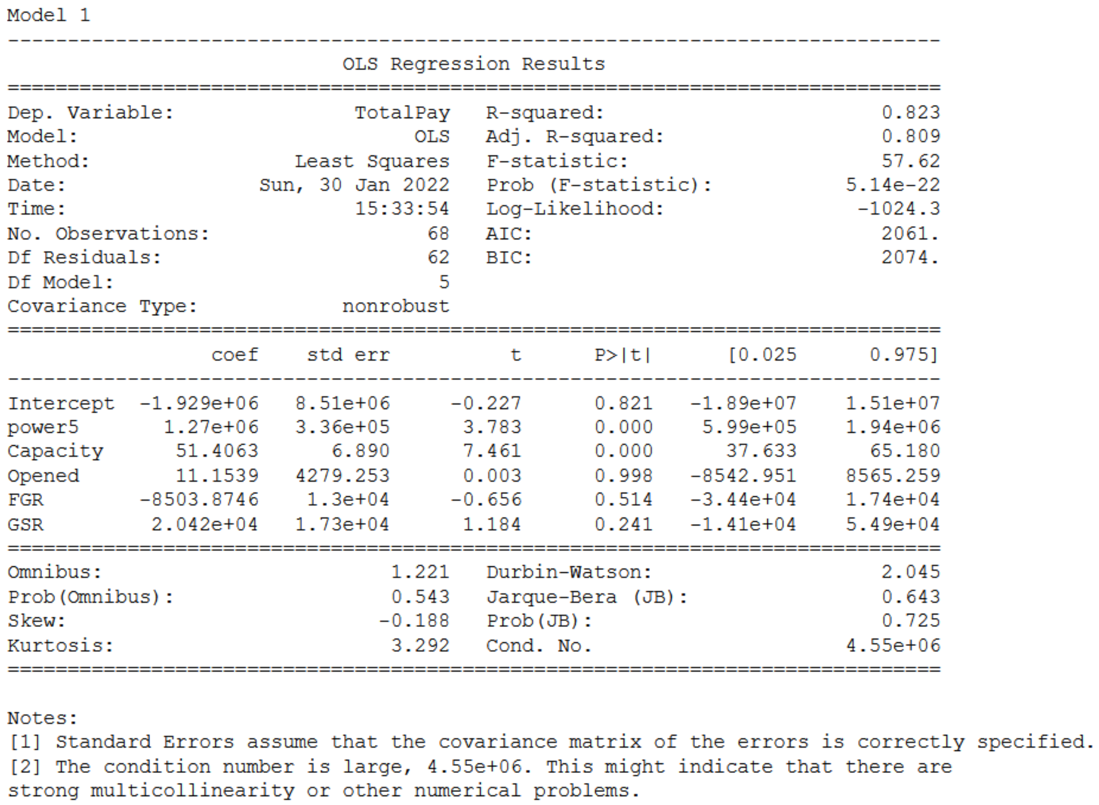


Figure 1.4

Linear Regression (Modeling)

Model 1 **(Figure 2.1)**

Based on simple visualization it’s clear that some variables would need to be removed/included in the initial model creation. The variables **‘power5 + Capacity + Opened + FGR + GSR’** were used and found that **Opened + FGR + GSR** don’t seem too significant with a p-value greater than .05. In cases of FGR and GSR, these two variable seem to have a correlation and possibly introducing the multicollinearity problem.



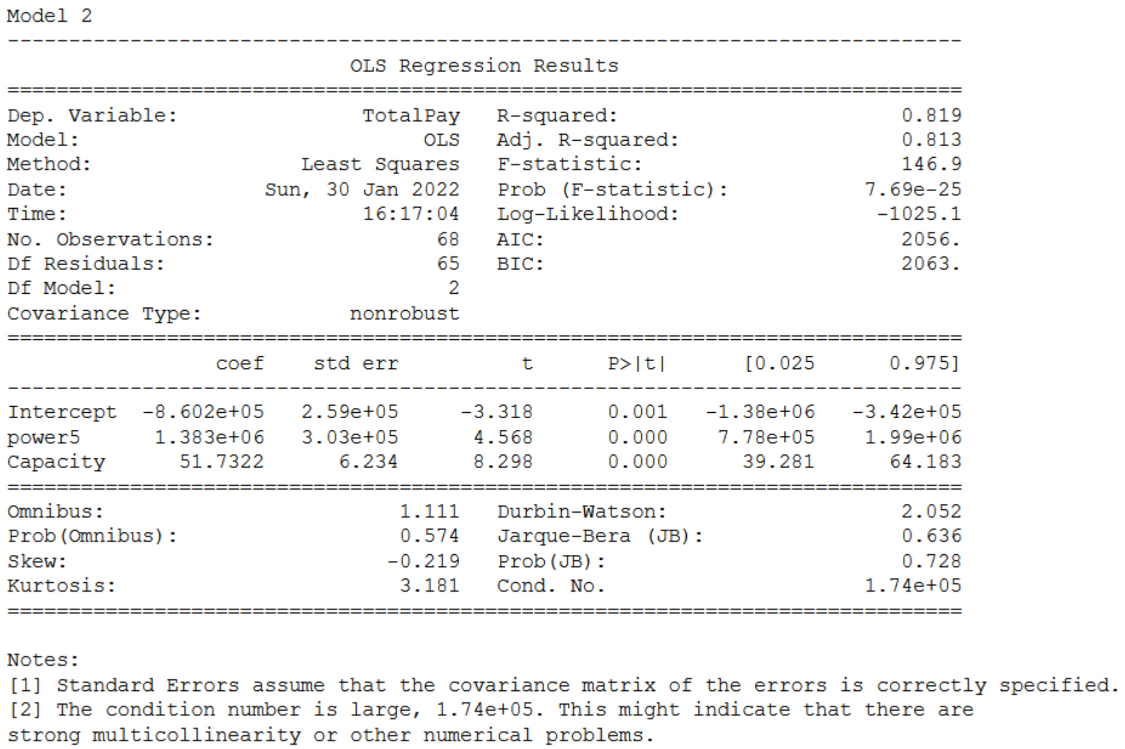
```
Model 1
------------------------------------------------------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:               TotalPay   R-squared:                       0.823
Model:                            OLS   Adj. R-squared:                  0.809
Method:                 Least Squares   F-statistic:                     57.62
Date:                Sun, 30 Jan 2022   Prob (F-statistic):           5.14e-22
Time:                        15:33:54   Log-Likelihood:                -1024.3
No. Observations:                  68   AIC:                             2061.
Df Residuals:                      62   BIC:                             2074.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -1.929e+06   8.51e+06     -0.227      0.821   -1.89e+07    1.51e+07
power5       1.27e+06   3.36e+05      3.783      0.000    5.99e+05    1.94e+06
Capacity      51.4063      6.890      7.461      0.000      37.633      65.180
Opened        11.1539   4279.253      0.003      0.998   -8542.951    8565.259
FGR        -8503.8746    1.3e+04     -0.656      0.514   -3.44e+04    1.74e+04
GSR         2.042e+04   1.73e+04      1.184      0.241   -1.41e+04    5.49e+04
==============================================================================
Omnibus:                        1.221   Durbin-Watson:                   2.045
Prob(Omnibus):                  0.543   Jarque-Bera (JB):                0.643
Skew:                          -0.188   Prob(JB):                        0.725
Kurtosis:                       3.292   Cond. No.                     4.55e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.55e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 2.1

Model 2 **(Figure 2.2)**

This model was created from Model 1 after removing the **Opened GSR,** and **FGR** variables because they were insignifcant, therefore it shouldn’t be something to include for predicting a coach’s salary. After switching to only keeping power5 and Capacity, the model had changed a lot (in a good way) with significant variables.



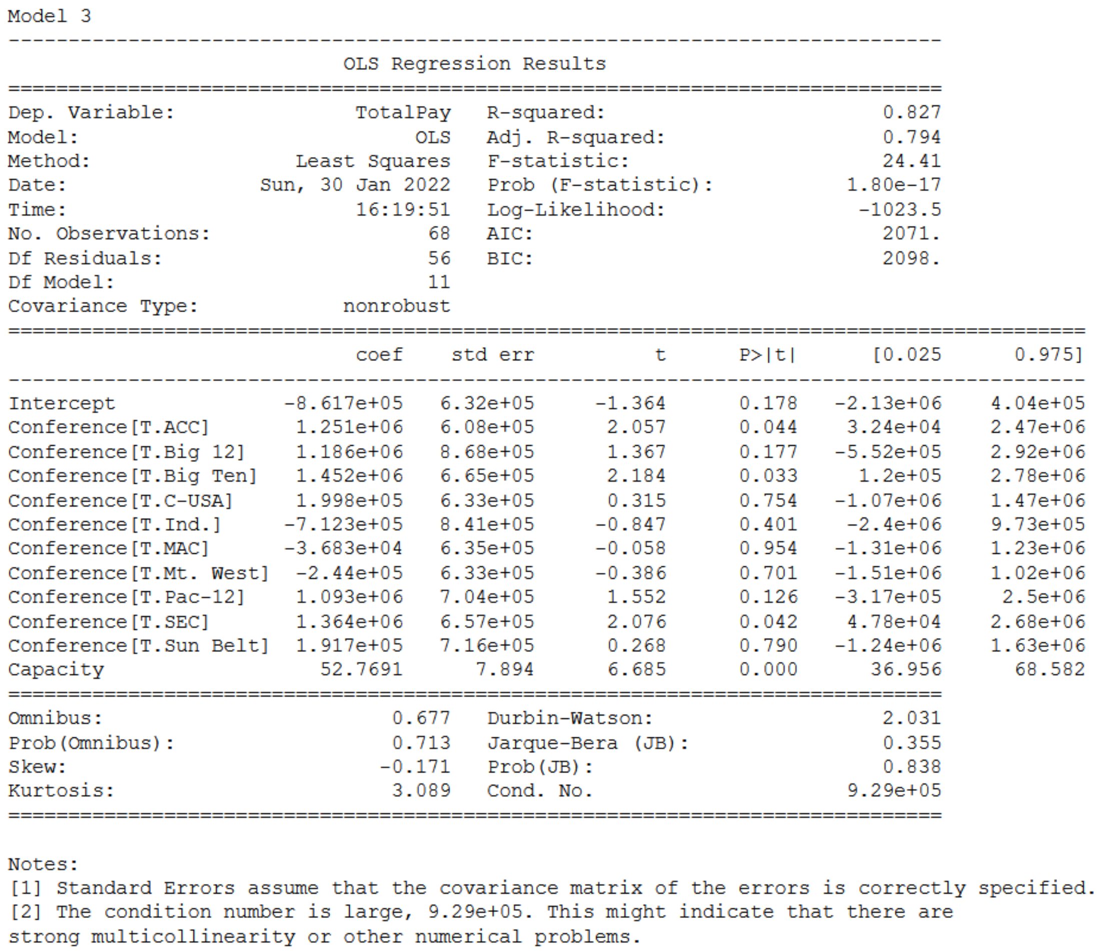
```
Model 2
------------------------------------------------------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:               TotalPay   R-squared:                       0.819
Model:                            OLS   Adj. R-squared:                  0.813
Method:                 Least Squares   F-statistic:                     146.9
Date:                Sun, 30 Jan 2022   Prob (F-statistic):           7.69e-25
Time:                        16:17:04   Log-Likelihood:                -1025.1
No. Observations:                  68   AIC:                             2056.
Df Residuals:                      65   BIC:                             2063.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -8.602e+05   2.59e+05     -3.318      0.001   -1.38e+06   -3.42e+05
power5      1.383e+06   3.03e+05      4.568      0.000    7.78e+05    1.99e+06
Capacity      51.7322      6.234      8.298      0.000      39.281      64.183
==============================================================================
Omnibus:                        1.111   Durbin-Watson:                   2.052
Prob(Omnibus):                  0.574   Jarque-Bera (JB):                0.636
Skew:                          -0.219   Prob(JB):                        0.728
Kurtosis:                       3.181   Cond. No.                     1.74e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.74e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 2.2

Model 3 **(Figure 2.3)**

Now there are only 2 variables that remain and since Conference was replaced with power5, perhaps switching it back could provide a different picture into what schools might not be as significant and also helps to answer some questions later on.



```
Model 3
==============================================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:               TotalPay   R-squared:                       0.827
Model:                            OLS   Adj. R-squared:                  0.794
Method:                 Least Squares   F-statistic:                     24.41
Date:                Sun, 30 Jan 2022   Prob (F-statistic):           1.80e-17
Time:                        16:19:51   Log-Likelihood:                -1023.5
No. Observations:                  68   AIC:                             2071.
Df Residuals:                      56   BIC:                             2098.
Df Model:                          11
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept              -8.617e+05   6.32e+05     -1.364      0.178   -2.13e+06    4.04e+05
Conference[T.ACC]       1.251e+06   6.08e+05      2.057      0.044    3.24e+04    2.47e+06
Conference[T.Big 12]    1.186e+06   8.68e+05      1.367      0.177   -5.52e+05    2.92e+06
Conference[T.Big Ten]   1.452e+06   6.65e+05      2.184      0.033     1.2e+05    2.78e+06
Conference[T.C-USA]     1.998e+05   6.33e+05      0.315      0.754   -1.07e+06    1.47e+06
Conference[T.Ind.]     -7.123e+05   8.41e+05     -0.847      0.401    -2.4e+06    9.73e+05
Conference[T.MAC]      -3.683e+04   6.35e+05     -0.058      0.954   -1.31e+06    1.23e+06
Conference[T.Mt. West]  -2.44e+05   6.33e+05     -0.386      0.701   -1.51e+06    1.02e+06
Conference[T.Pac-12]    1.093e+06   7.04e+05      1.552      0.126   -3.17e+05     2.5e+06
Conference[T.SEC]       1.364e+06   6.57e+05      2.076      0.042    4.78e+04    2.68e+06
Conference[T.Sun Belt]  1.917e+05   7.16e+05      0.268      0.790   -1.24e+06    1.63e+06
Capacity                  52.7691      7.894      6.685      0.000      36.956      68.582
==============================================================================
Omnibus:                        0.677   Durbin-Watson:                   2.031
Prob(Omnibus):                  0.713   Jarque-Bera (JB):                0.355
Skew:                          -0.171   Prob(JB):                        0.838
Kurtosis:                       3.089   Cond. No.                     9.29e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.29e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 2.3

Questions

What is the recommended salary for the Syracuse football coach?

- The recommended salary range would be $3,002,165 to $3,084,650

What would his salary be if we were still in the Big East? What if we went to the Big Ten?

- Big East (aka AAC): $1,701,585 to $1,751,476
- Big Ten: $3,084,650 to $3,203,270

What schools did we drop from our data and why?

```
Schools removed due to inconsistency in naming when joining, contained
missing data, or outlier:
--------------------------------------------------------------------------
1 Air Force
2 Alabama
```

```
3 Alabama at Birmingham
4 Arizona State
5 Arkansas State
6 Army
7 Buffalo
8 Connecticut
9 Florida International
10 Kansas State
11 Liberty
12 Louisiana-Lafayette
13 Louisiana-Monroe
14 LSU
15 Massachusetts
16 Miami (Ohio)
17 Michigan State
18 Mississippi
19 Navy
20 Nevada-Las Vegas
21 North Carolina State
22 Northern Illinois
23 Oklahoma State
24 Oregon State
25 Southern California
26 Texas
27 Texas Christian
28 Texas-El Paso
29 Texas-San Antonio
30 UCLA
31 Virginia Tech
32 Washington State
33 Baylor
34 Brigham Young
35 Rice
36 Southern Methodist
```

What effect does graduation rate have on the projected salary?

- It doesn't have any effect. This is an insignificant variable that was removed since it only creates additional noise

How good is our model?

- It does a pretty good job of accounting for most of the data points and predicting the recommended salary. The best model here (Model 3) has an R-squared of 0.827

What is the single biggest impact on salary size?

- Based on the breakdown between the most significant models, the remaining variables are **Capacity** and **(Conference or power5)**. It is clear that which variable would alter the salary size greatly and that is **Conference** because it can be seen above that if Syracuse was in a different conference the Coach's salary would shift very significantly.