Alex Xushao Yu
Kyle Neville
Matt Turk
Max Blane

Multiple Linear Regression to predict Player Wages

Abstract

In almost all cases, the best performing athletes are the ones who make the most money in their given sport. But what are the criteria for the best players, and how are they assessed? In fact, there are many factors that determine how much any singular player is paid. The purpose of this project was to create a method of determining a soccer player's salary based on the information given in FIFA's database. In order to develop a successful model for the prediction of a FIFA player's salary, we needed to apply several methods of data analysis.

We began with a multiple linear regression model that had only a few predictors, namely club, overall, potential, and international reputation. Then, using the two data sets (training and testing) for the basis of our analysis, we studied the effects of response and predictor transformation, creating a bunch of new variables along the way (max, tier, unique, avg, WR, OP, etc) but ultimately through AIC and BIC testing, as well as significance tests and other statistical modeling methods, we discarded these predictors and employed others instead, to be discussed later in this paper. Over time, we were able to improve the effectiveness of our model, while not overcomplicating it. We imputed missing values, created new variables, and crafted interaction terms in order to make our data set more useful. Our final model had seven predictors and yielded an r-squared value of 0.912 on the training data.. Our highest r-squared value submitted to Kaggle was 0.9553 (private score), ranking tenth on the Kaggle Public Leaderboard and nineteenth on the Kaggle private leaderboard.

Introduction

Our goal was to develop a multiple linear regression model to best predict a FIFA player's salary. We developed this model using the latest (2019) data from FIFA's database, including observations of over 18,000 FIFA players. Each observation contained 79 variables with information about the player including their specific skills, general athleticism, body type, nationality, and more. While it is probable that many of these factors play a significant role in determining a FIFA player's salary, as statisticians, our aim was to find a small number of the most significant variables that could best predict a player's salary in order to have a simple, valid regression model. Our final FIFA Training Data Set included 11640 rows and 97 columns, including 18 new columns, some of which were used in the final model.

*The Work.Rate column was lost along the way in the imputation, but the predictor wasn't intended to be used in the final model anyway, and so the error was of little significance

Methodology

Variable Selection and New Variables (Included in FINAL MODEL only)

To decide which variables to use for the regression, we began by looking at the matrix and correlation plots of the several different variables, taking note of their relationship with WageNew. Some of the most significant predictors were categorical variables that had too many categories to include in our model. To use these predictors we created new variables to capture the important information that these predictors give without using all of their categories.

The first, and most significant of these was the "club" variable. To break apart the club variable, we created 4 new numerical variables: ClubMaxWage, ClubMinWage, ClubMeanWage, and ClubMeanWageDensity. Each of these were created in a similar manner. For example, ClubMeanWage was created as follows: first, we grouped the players by club, and then found the mean wage for that club. Next, we created a new variable called ClubMeanWage and assigned each player the mean wage calculated for their club. This same procedure was done with maximum wage, minimum wage, and club wage density. Because each of these variables was created grouping by club, they capture information built into the club variable, without explicitly using it in the model.

We also wanted to use the players positions as predictors but ran into a similar problem as above. To make the model more efficient, we created a new variable called position group. This was simply made by grouping similar positions. We turned 24 positions into 10 position groups to use as a predictor.

A third variable we created called AbilityProfile was created to better utilize player attributes such as stamina, crossing, ball control, etc. We realized that the importance of these attributes depends on the position of the player (for example, reactions are more important for GK than for CM). For this new variable, a fixed set of important players attributes has been assigned to specific position groups. Each player's ability profile is the normalized sum of their associated attribute scores. Each score is out of 100, so a player's ability profile is a number between 0 and 1, with 1 representing a perfect score on each attribute.

Finally, after seeing that players in a particular age range tended to make higher salaries than those in others, we created a variable called "old" that groups the player into three different age

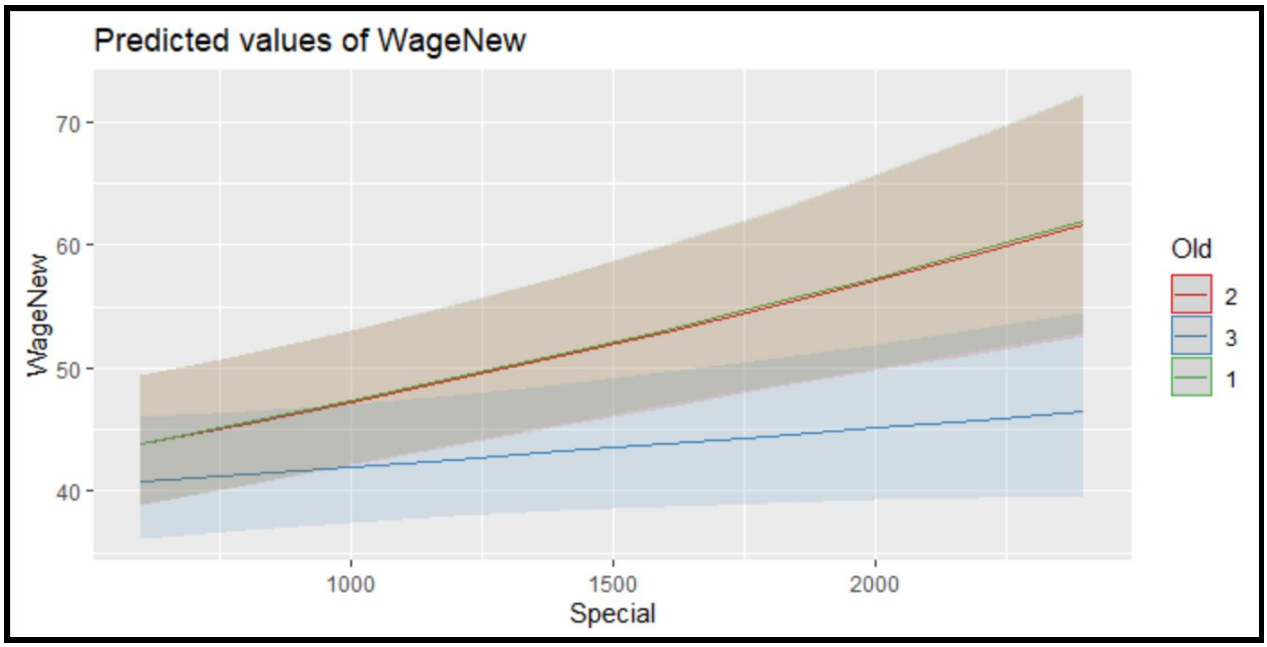
groups (less than 24, between 24 and 28, and greater than 28). With each of these new variables, we were able to apply key insights into our model to simplify and improve it.

Transformation

We performed a power transform on these variables which concluded that the correct transformation was to take the log of WageNew, ClubMinWage, and ClubMaxWage. The additional suggestions from the power transform resulted in what was overfitting of the response variable, and so, these were the only transformations used.

Interaction Plots

Furthermore, we chose to include a few interaction terms in our model as demonstrated by the following interaction plots:

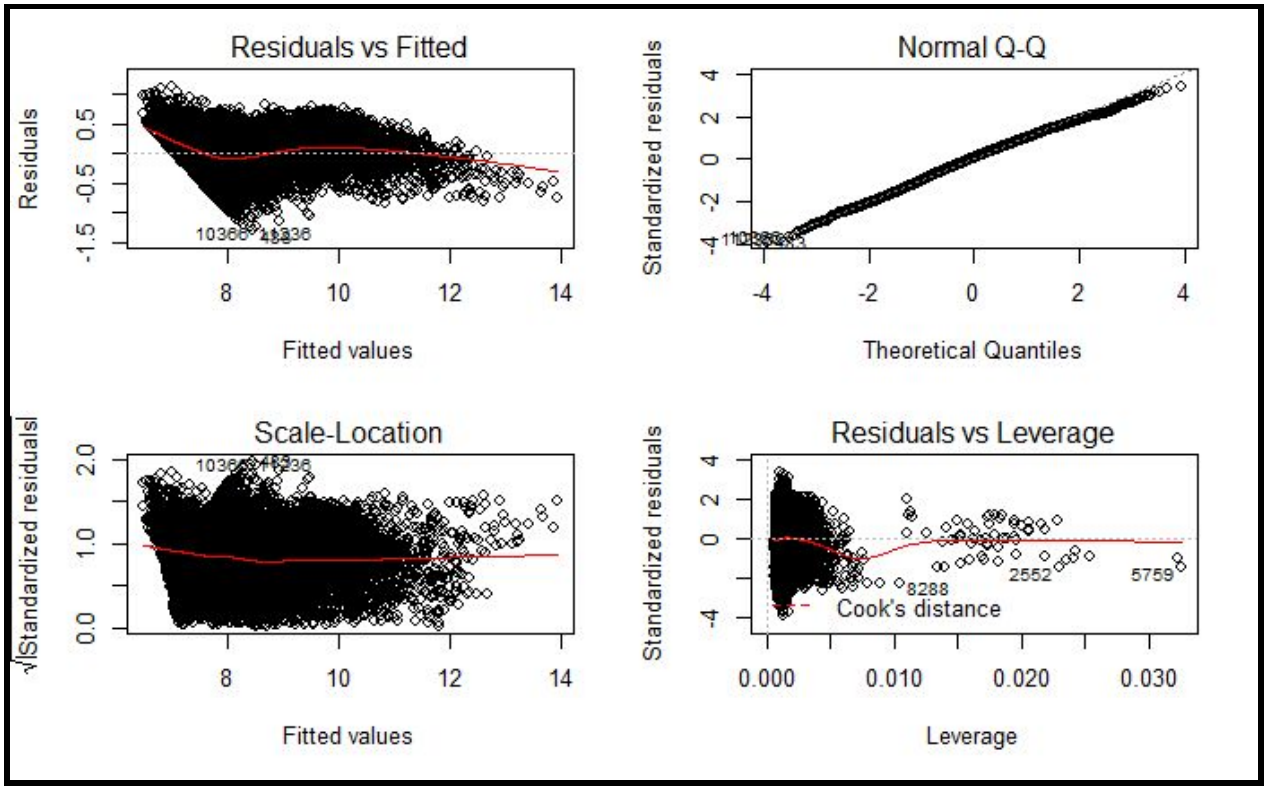


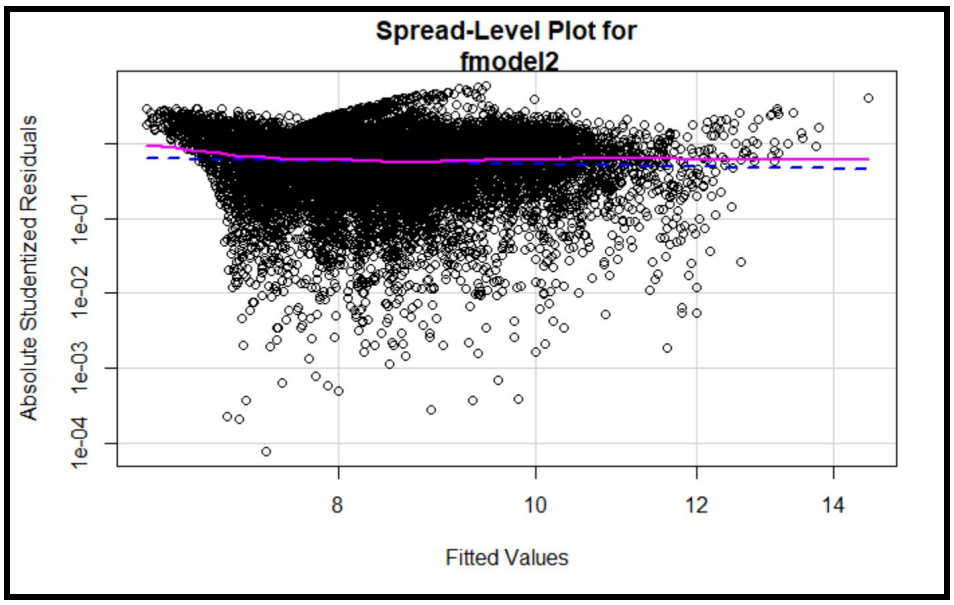
The Model

This selection led us to the following model:

log(WageNew) ~ log(ClubMinWage) + log(ClubMaxWage) + ClubMeanWageDensity + PositionGroup:AbilityProfile + Overall:AbilityProfile + Special:Old + Special:AbilityProfile

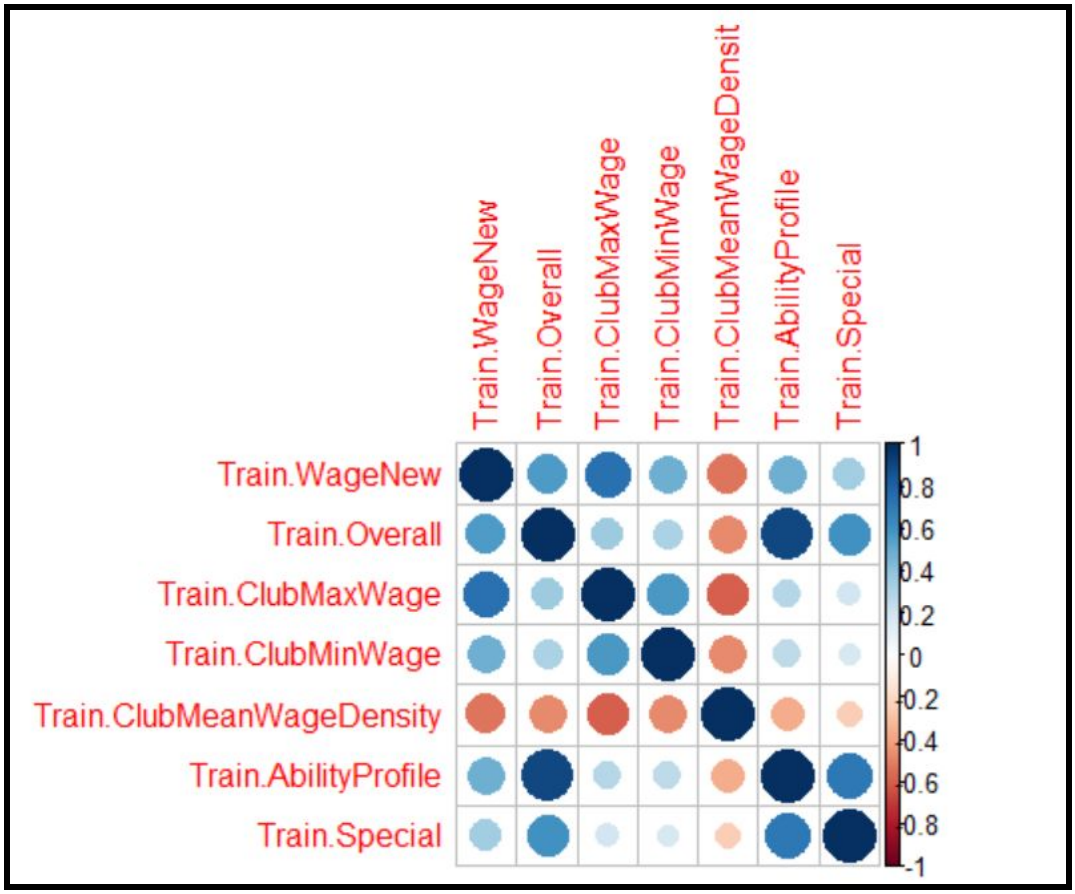
Diagnostic Plots





Correlation Matrix

The following plot displays correlation of the predictors in the model, excluding interaction terms, and are labeled as follows:

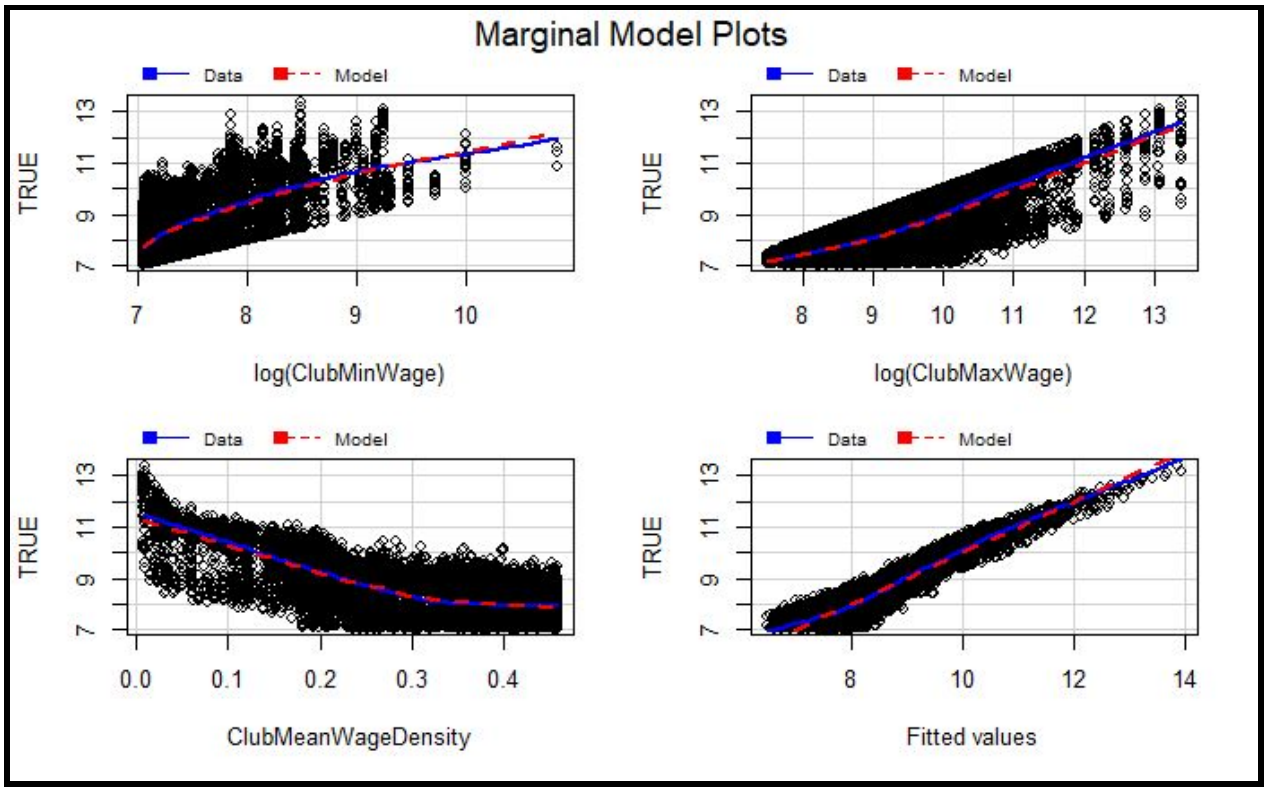


VIF

The following VIF's were calculated using the model's predictors *without* the interaction terms:

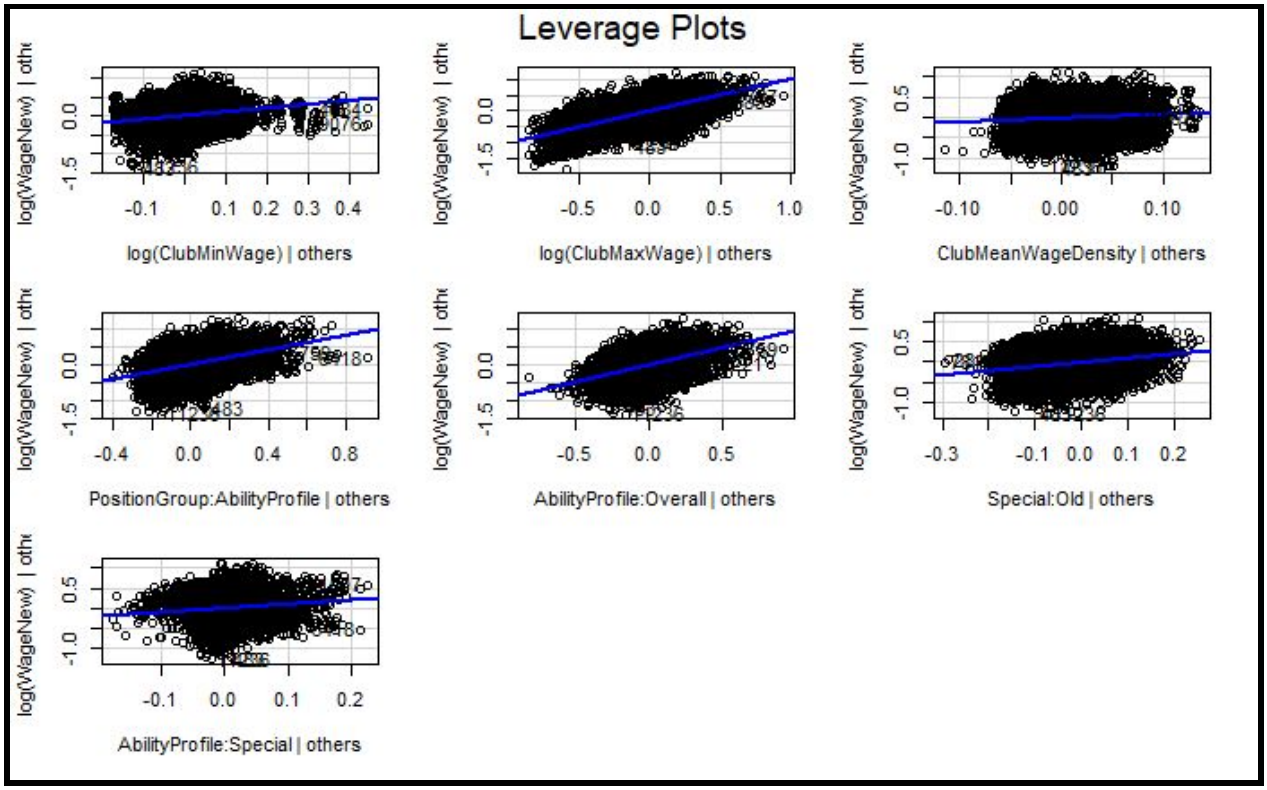
Predictor	**VIF**
log(ClubMinWage)	1.688442
ClubMeanWageDensity	1.721765
log(ClubMaxWage)	1.794372
PositionGroup	1.139988
AbilityProfile	3.391025
Overall	3.457686
Special	3.215260
Old	1.090583

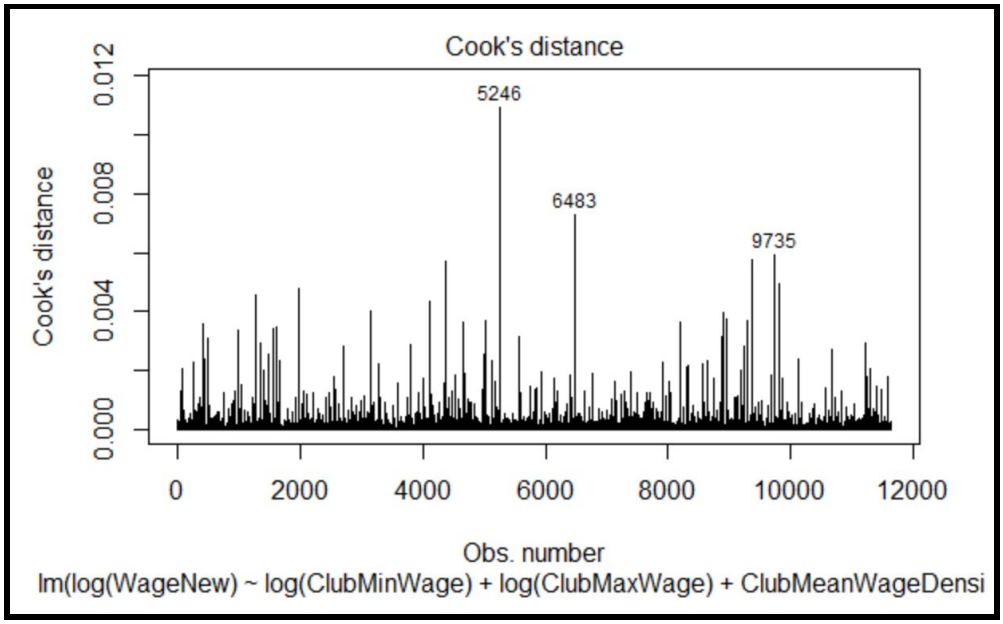
Marginal Model Plots



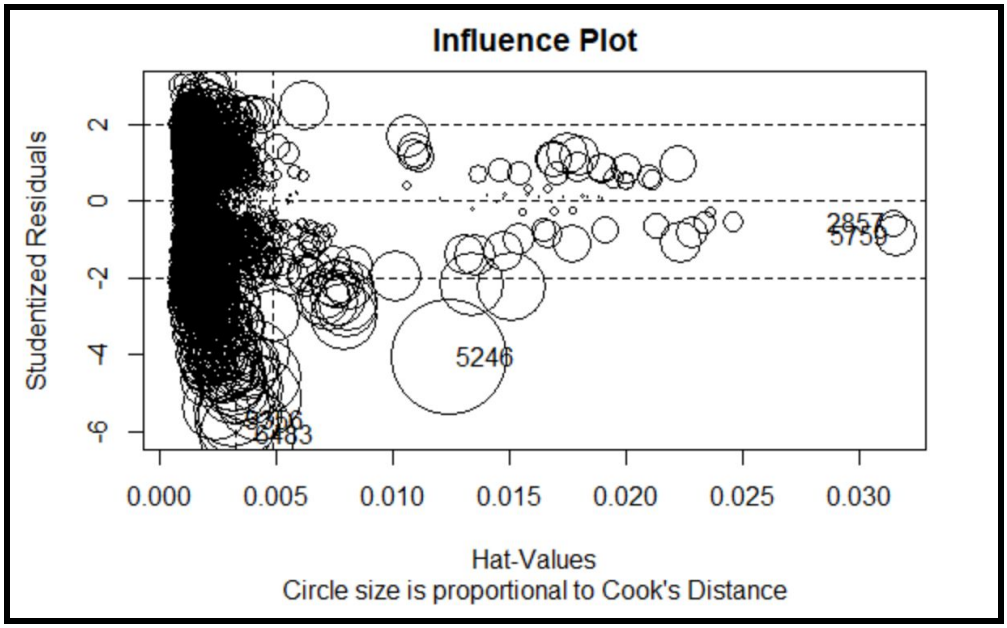
Leverage and Leverage Points

The following plots, on the below page, show 1) whether the predictors, and interaction terms are good predictors on their own or if they rely on the other predictors as well as 2) which particular points serve as significant good/bad leverage points in the updated model.





This plot shows that points 5246, 6485, and 9735 are bad leverage points.



AIC and BIC Test

Both AIC and BIC methods recommended our full model, run both forward and backward.

```
Start:  AIC=-22840.66
log(WageNew) ~ log(ClubMinWage) + log(ClubMaxWage) + ClubMeanWageDensity +
    PositionGroup:AbilityProfile + Overall:AbilityProfile + Special:Old +
    Special:AbilityProfile

                                Df Sum of Sq    RSS    AIC
<none>                                       1610.8 -22841
- ClubMeanWageDensity            1      2.18 1613.0 -22834
- AbilityProfile:Special         1      8.23 1619.1 -22791
- log(ClubMinWage)               1     45.27 1656.1 -22528
- Special:Old                    3     83.67 1694.5 -22280
- PositionGroup:AbilityProfile 10    171.50 1782.3 -21758
- AbilityProfile:Overall         1    279.63 1890.5 -20987
- log(ClubMaxWage)               1    896.76 2507.6 -17699
```

Results and Discussion

As has been seen, our model log(WageNew) ~ log(ClubMinWage) + log(ClubMaxWage) + ClubMeanWageDensity + PositionGroup:AbilityProfile + Overall:AbilityProfile + Special:Old + Special:AbilityProfile has a 0.912 R^2 on the testing data, with all of the VIF values are under 5,

indicating that there is not a multicollinearity problem. In addition, our final model had a BIC score of [19.00, -25069.49].

Our R output indicated that all of our predictors were significant both as coefficients in the summary output, and in the anova output as well. The residuals are normally distributed as is shown in the QQ_plot above, and the Residuals vs Fitted plot and Residuals vs. Leverage plots show that our model maintains constant variance. From the Marginal Model plots, we can see that the predictor model we created successfully follows the smooth curve of the response variable, and thus no further transformations need to be made to the model. Also, the leverage plots confirm the results that we saw in our vif test, that all of the plots contain trend lines (are not flat) and therefore do not have too significant of a dependence on the other predictors in the model. Lastly, after taking out with the tests, the bad leverage points that R could identify, the cook's distance showed that the overall model was not considerably affected by the leverage points, and that only points 5246, 6485 and 9735 were bad leverage points.

The most important part for this project was trying to get the highest adjusted R-squared possible while keeping the VIF-value under the reasonable level of 5 all the while maintaining valid diagnostic plots. The most difficult challenge we faced was improving the model's R-squared value while not violating assumptions of linearity, normality, little to no multicollinearity, and homoscedasticity. It seemed as though every time we would optimize one side, the other side would falter. However, ultimately when we came to what would serve as our final model, we saw that despite a few percentage points drop in R^2, based on the diagnostic plots, VIF, marginal model plots, etc., our final model met all of the conditions and assumptions for being a valid multiple linear regression model, and we thus decided to keep it as it was.

Limitations and Conclusions

One of our most significant challenges was incorporating the Club predictor into the model. After learning that Club was the best single predictor of salary, we also learned that it was improper to include it in the model because it had too many categories and thus made the model too complex. To overcome this issue, we initially organized clubs into tiers based on the average salary for each club. This new club wage tier categorical variable had high correlation with mean club wage, but we found that it still did not help the model as much as we expected. We then discovered that splitting Club into two predictors, ClubMinWage and ClubMaxWage (which indicated the minimum and maximum wages for each club, respectively) was a better way to represent Club with a dummy variable, but still not perfect. So, while we could not use our optimal model with the Club variable, we found an appropriate substitution while importantly maintaining a valid model. In addition, the Nationality column looked to be a good predictor for Wage, but because the Test and Train data set included differing levels/factors of Nationality, the

variable could not be used, as we didn't have enough time to edit the variable in order for it to work and be included in the final model. Lastly, while we decided to employ a linear model to solve this problem, it is possible that some other kind of model would have been more effective.

References

Sheather, Simon J. *A Modern Approach to Regression with R.* Springer New York, 2009.