Report on MLS Player Annual Earning Linear Regression Model

Introduction:

The high-salary phenomenon in the sports field attracts much attention from society. By fitting a linear regression model on various statistics of Major League Soccer(MLS) players and their salaries in 2018, I would like to answer my research question: What is the relationship between the annual earning of an MLS player and his performance in the year?

The importance of this research is that it is beneficial for the soccer market. Using the linear model, soccer clubs can better estimate a player's value. Also, players can have a direction for improvement to increase their earnings.

Methods:

Salary was the average annual compensation of an MLS player, and it was the initial response in my model. *GP, GS, G, A, SHTS, SOG, FS, OFF* were different statistics of a player's performance.

Model Validation:

I randomly divided my dataset into a training set with 60%, and a validation dataset with 40% of the original dataset. Thus, I had enough data to generate a linear model, while the validation dataset contained similar data. I found my final model based on the training dataset and fitted it on the validation dataset. I tried to validate my model by checking if the model produced similar estimated coefficients, significant predictors, R^2 and adjusted R^2. I also checked if new model violations and multicollinearity appeared. If the two models were similar, my model was validated.

Model Violations:

I first performed an exploratory data analysis(EDA) on the training set. Thus, potential issues such as assumption violations, multicollinearity, and influential points were anticipated.

The initial model contained all possible predictors without transformations. I checked condition 1 by plotting the fitted value versus the response value. If points were randomly scattered around the identity function, I argued that condition 1 was satisfied. If no plot showed a non-linear relationship between predictors, I argued that condition 2 was satisfied. When conditions are satisfied, I plotted the residual plots for every variable and a QQ plot. I checked if the residuals are randomly scattered around 0 without a non-linear pattern, and if the QQ plot does not have a large deviation at the end. Combining all of my observations, I argued if assumptions were violated. To fix assumption violations while maintaining the interpretability, I used simple transformations such as logarithmic transformation referenced by box-cox transformation. I also used researches and context to argue the removal of problematic variables.

Variable Selection:

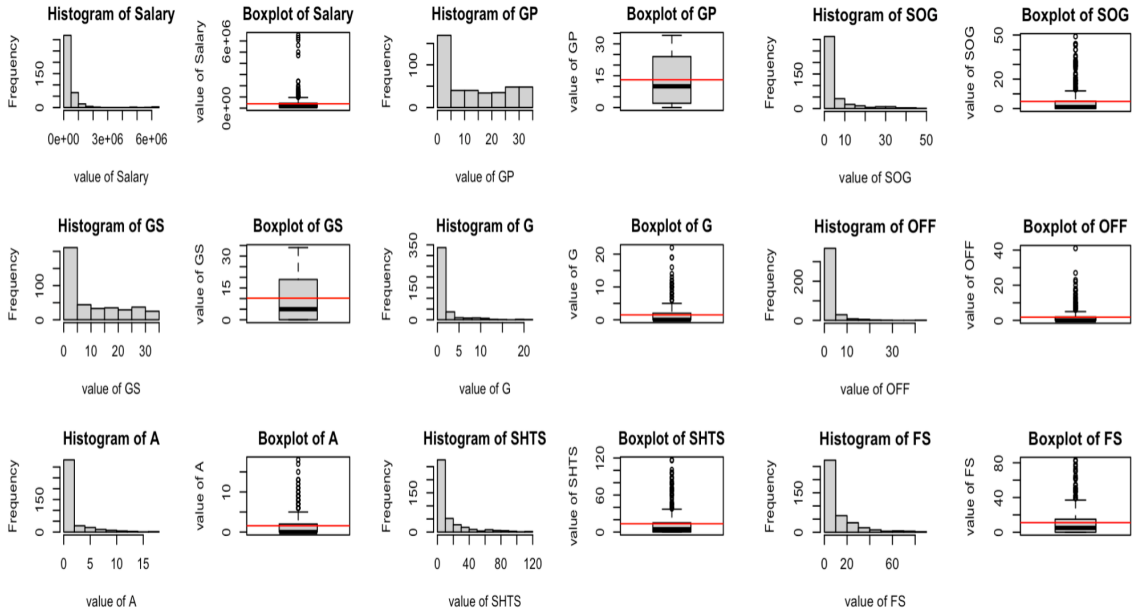
With a model that satisfied all assumptions, I checked the multicollinearity by calculating variance inflation factors(VIF) for each predictor. If the VIF exceeds 5, the predictor had severe multicollinearity. They could increase standard errors and make coefficients the wrong sign, so I removed some of them based on researches, context, and significance of variables. I made sure no new assumption violations appeared when problematic predictors were removed.

I used stepwise selection based on Bayesian Information Criterion(BIC) because it accounted for the conditional nature of the model comparing to the forward and backward selection. I applied it

for both sets of possible predictors with or without multicollinearity removed.
For the two resulting models, I checked assumptions and multicollinearity again to see if they were concerning. I chose a model with high R^2, small BIC, and without assumption violation and multicollinearity as my final model. I used a partial F test to confirm that other predictors should not be in the model.
After that, I detected leverage points with leverage higher than the cut-off and outliers with standardized residuals higher than the cut-off. I also found influential points using the cook's distance, DEFECTS, DFBETAs. They are problematic as they could influence the placement of the regression line. By EDA there was no measuring error, so no observation should be removed.

Results:

I found no missing information in my training dataset. Potential predictors all had many zero values, so I should be cautious when applying transformations. All variables were highly right-skewed, then the assumption of normality was probably violated, and transformations were needed. There were also many outliers for variables except for GP and GS, which means there might be problematic observations in the resulting model.



Distributions of Variables

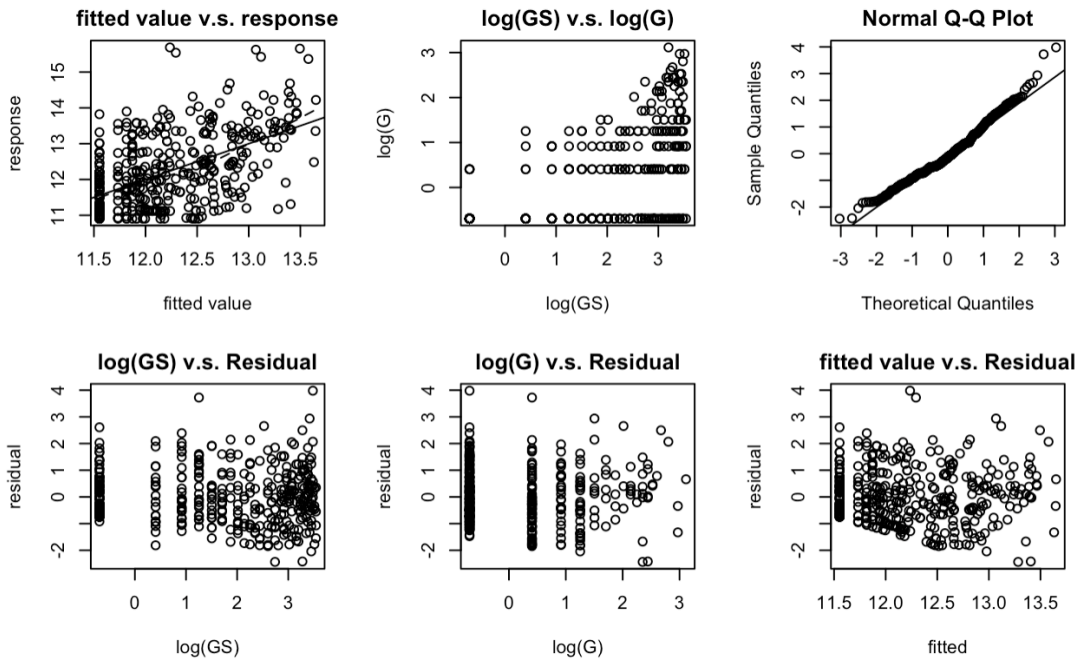
For the initial model, it was enough to say conditions 1 and 2 were satisfied. The residual plots were all fine, but there was a downward trend in the residual versus fitted values plot, and the QQ plot went very upward at the end. Thus, I applied a log transformation on all variables. Since predictors had many zero values, I added 0.5 to the observations of all predictors. The new model satisfied conditions 1 and 2 well and showed a good QQ plot and residual plots. I concluded that this model satisfied the linear regression assumptions, which meant I could use the result from this model to make decisions.
I calculated VIF for all the predictors and found that log(GP), log(GS), log(SHTS), log(SOG),

log(FS) had VIF >5. From the correlation plot in the appendix, I knew that log(GP) and log(GS) were positively linearly related. Since GS, the number of games the player started, was previously shown to be a significant factor for wage(Celik & Ince-Yenilmez, 2017), I did not remove it. Instead, I removed GP, the number of games the player played, which had a similar definition as GS. Other predictors were not significant in the transformed model, so it was reasonable for me to remove them to resolve multicollinearity.
I applied stepwise selection methods based on BIC for both sets of transformed predictors with or without multicollinearity removed, and ended up with two potential models.

	Model 1	**Model 2**
Initial Multicollinearity	Not Removed	Removed
Model	log(Salary)~log(G)+log(GS)+log(GP)	log(Salary)~log(G)+log(GS)
Final Multicollinearity	log(GS), log(GP)	None
BIC	-99.09	-96.79
R^2	0.318	0.3041
Adjusted R^2	0.313	0.301
Multicollinearity	None	None

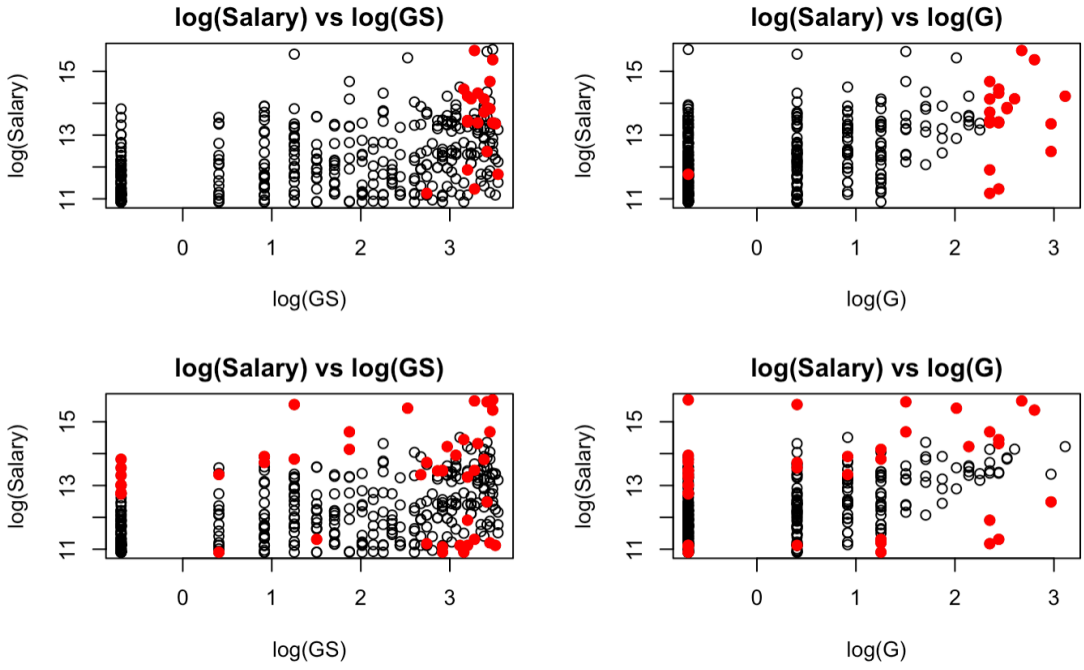
Comparing Models Produced by Automated Selection

I performed partial F tests on transformed model versus Model 1, reduced model versus Model 2, and Model 1 versus Model 2. Only the last one showed a significant p-value, meaning that only log(G), log(GS), log(GP) should remain in the model.
Since Model 1 was Model 2 with an extra predictor, I compared their adjusted R^2. Model 2's adjusted R^2 and R^2 were only 0.01 unit lower than Model 1.
Although Model 1 had a higher BIC, it had severe multicollinearity on log(GP) and log(GS). Overall, I argued that Model 2 was more preferable and was my final model.
Goodness of Final Model:
Based on the plot of fitted value versus response, Model 2 satisfied condition 1. Although the graph of log(G) versus log(GS) had some fanning pattern, it might be due to the zeros in both variables. Considering the points were also scattered, it was enough to say condition 2 was satisfied. The QQ plot and residual plots looked fine, so I argued that the final model satisfies the assumptions of linear regression models.



Condition 1 & 2, Residual Plots and QQ Plot of Final Model

The model had no outliers, but some leverage points are shown in the first row of the following plots. There were also a few influential points, shown in the second row.



Leverage Points(1st Row), Influential Points(2nd Row)

Final Model: log(Salary) ~ log(GS) + log(G)					
	Training Dataset			**Validation Dataset**	
	Estimate	Std. Error	Significance	Estimate	Significance
Intercept	11.9	0.07	<0.001	12	<0.001
log(GS)	0.16	0.03	<0.001	0.12	<0.05
log(G)	0.38	0.05	<0.001	0.33	<0.001
Residual standard error	0.8736			0.9301	
R^2	0.3041			0.219	
Adjusted R^2	0.30			0.2133	
Multicollinearity	None			None	
Leverage Points	21			16	
Outliers	None			None	
Influential Points	46			26	

Comparing Training Model and Validation Model

Validation of Final Model:

For the model fitted on the validation dataset, all estimated coefficients were still significant and within 2 standard errors from the training model coefficients. It also had a similar residual standard error and number of leverage points and outliers. As shown in the Appendix, no new assumption violations, multicollinearity appeared in the validation model. There were fewer influential points which were reasonable since the dataset was smaller. Although the adjusted R^2 decreased by 0.9, most values corresponded to the training model. Overall, there was enough evidence to validate my model.

Discussion:

The final model was: log(Salary) ~ log(GS) + log(G), with the estimated coefficients presented in the table above. GS means number of games the player started and G means number of goals the player scored.

For example, it showed that when the logarithm of the number of goals the player scored increased by 1, the average logarithmic of the annual earning of the player increased by 0.38. The result from the model was valuable because it suggested players to practice scoring which had a significant positive impact on their salary. Soccer clubs knew that players in the starting lineup should have a higher salary.

Limitations:

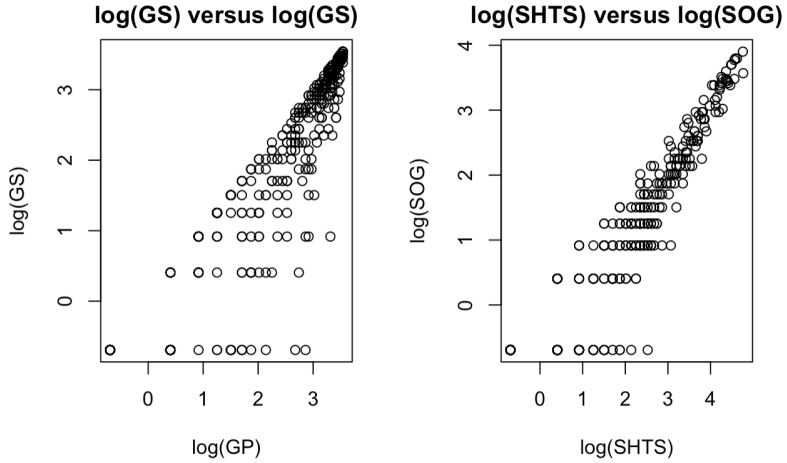
Although the data was reliable since they are from official websites, sampling bias might occur if observations were lost during the scraping process.

Since the model only had two predictors, it might be under-fitting, and some predictors in the true relationship might not be included. They could be confounding variables that were hard to obtain, such as players' popularity, or predictors eliminated by researches and context at the beginning. Missing them affects the generalizability of the model to a larger population.
There were many leverage and influential points. They had a disproportionate impact on the estimated regression line and made it less accurate. The automated selection did not know the impact of these observations, and did not check assumptions during the selection process. Thus, it might not choose the optimal solution. Also, it might be biased because we did not create a hypothesis but searched for the most significant results.

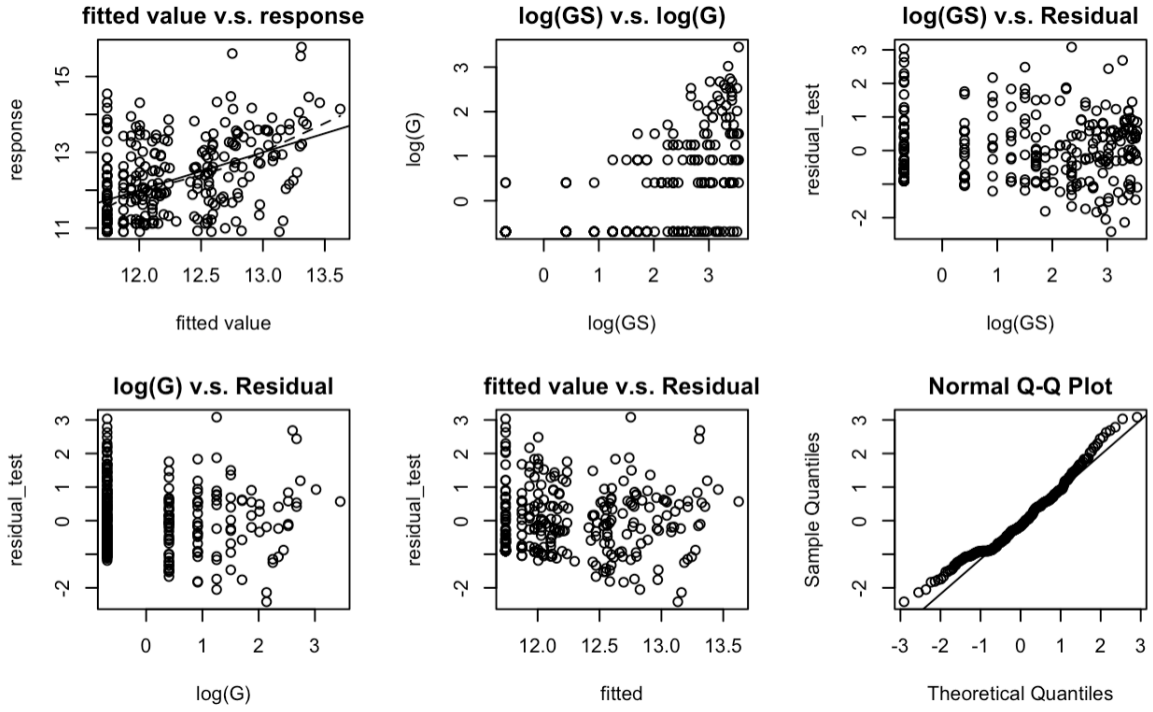
References

Celik, O. B., & Ince-Yenilmez, M. (2017). Salary differences under the salary cap in Major League Soccer. *International Journal of Sports Science & Coaching*, *12*(5), 623–634. https://doi.org/10.1177/1747954117727809

Appendix



Linearity Observed in Pairs of Predictors



Condition 1 & 2, Residual Plots, QQ plot for Final Model on Test Dataset