University of Toronto
STA302 - Methods of Data Analysis I
Final Project

Using Multiple Linear Regression to Predict Salaries of NBA Players Based on Game Statistics

Presented By:

Richard Li
Matthew Zannier
Abdullah Motasim
Ahmed Hassini
Malek Jaziri

1.0 Introduction

The research question our team decided to answer was what factor has the most effect on an NBA player's salary. We are looking at a broad range of statistics such games started in (GS), minutes played (MP), number of 2 or 3 point basket made (FG), number of 2 or 3 point basket attempted (FGA), number of 3 point shots made(X3P), number of 3 point shots attempted (X3PA), number of 2 point shots made(X2P), number of 2 point shots attempted (X2PA), number of free throws made (FT), number of free throws attempted (FTA), number of defensive rebounds (DRB), number of assists (AST), number of turnovers (TOV), number of points scored (PTS). Note all these statistics are recorded per game.

This research question interested our group as all of us have an interest in basketball and wanted to know what the most sought-after statistic was for players at a professional level. The results of this question might be of interest to people such as NBA players to determine what statistic they should improve to increase the amount they are getting paid. Also, the results might be of interest to recruits bringing new talent into the NBA to determine how much a potential player would cost their team based on their college stats.

Conducting online research we have found this is a well-researched topic with many papers trying to answer what statistics contribute to a player's salary. *The Sport Journal* has a research paper that uses multiple regression to determine a relationship between a player's salary and career performance variables (Lyons Jr. et al.). Some important variables they used were offensive and defensive statistics which we will also try to utilize. Also, an important result from their research was that points per game and field goal percentage were the two main contributors to the player's salary, this result is further supported by another paper which found points per game were a significant indicator of a player's salary (*Schmidt*). Also, we would like to note that although the papers we found answer the same question as us we may find completely different results as these papers look at data 10 years old and the NBA has changed over the years causing different statistics to become more desired.

2.0 Methods

This section will outline the methods, tools, and techniques we will use to derive our final model.

2.1 Variable Selection

Our data was gathered from *Basketball-Reference* for the 2023-2024 year NBA season (NBA). As mentioned in the introduction there are many variables we could take into consideration when trying to determine an NBA player's salary, but many of them such as age or steals don't have a linear relationship. To mitigate this issue we will calculate the correlation matrix between each of our variables. This will help us to identify the strength and direction between each pair of variables. Furthermore, we can also examine the correlation between each predictor and response to focus on only the relevant predictors ensuring our model has high efficiency and effectiveness. We have decided to use two main techniques: Akaike Information Criterion and Stepwise Analysis paired with our correlation matrix to determine our variables.

2.1.1 Akaike Information Criterion (AIC)

The AIC will be used to measure the relative quality of our statistical models and aid us in model selection. The AIC is calculated using the following formula:
$AIC = 2k - 2\ln(L)$ where k represents the number of parameters in the model and L represents the likelihood of the model.

The reason we chose to use AIC is because it will create a balance between model fit and complexity. Lower AIC values mean the model is a better fit, however, AIC also considers the number of parameters in order to prevent overfitting. This is crucial for us as we have 26 possible predictors and we must ensure our model uses the least amount of predictors while still maintaining high accuracy.

2.1.2 Stepwise Analysis

Stepwise analysis is used to identify the most relevant subset of predictors in a model with a large set of predictors. In our case we will use stepwise analysis in order to narrow down the 26 predictors to only select the ones that contribute significantly to explaining the variation in our response, salary. There are two main types of selection: forwards selection and backwards selection.

Forward selection will be used to determine whether we must add predictors. Forward selection process is adding predictors one at a time with the best change in AIC. This process stops when no more predictors improve the model's AIC. While backward selection is similar to forward except it is used to determine if predictors are able to be removed. Backward selection process is to continuously remove predictors one at a time that least affect the model's AIC. This process stops when removing additional predictors results in a significantly worse AIC. For our selection process we will utilize the benefits of both forward and backwards selection.

2.1.3 Model Creation Steps

Due to the complexity of our data we decided to pick the best 4 predictors with the highest correlation between itself and the salary. Next, we created another model with all the predictors with a correlation between itself and the salary over 0.5. We then applied Stepwise analysis utilizing the AIC as the selection metric to both models, ultimately resulting in 4 distinct models. Please see section 3.2 for further details.

2.2 Model Validation

To validate the model, we will randomly split our dataset into two parts: a training dataset and a test dataset with a 75/25 split, respectively. We will use the training dataset to perform all model building and diagnostics until we reach a final model, and then we will use the test dataset to evaluate its effectiveness.

2.2.1 Training Dataset

Using our training dataset, we first need to perform several analyses to ensure it is ready to be used for creating our model. This involves performing an exploratory analysis, verifying that the assumptions hold, dealing with problematic observations, checking for collinearity, and performing model selection as described in the variables selection section.

2.2.2 Exploratory Analysis and Assumption Checks

By utilizing exploratory analysis, we can quickly check and affirm the necessary assumptions:

Linearity: We assess the linearity assumption by plotting each predictor against the response variable. We will look for a general linear pattern in the plots.

Normality of Residuals: We affirm the normality of residuals by using a Q-Q plot. We will check that the values closely follow the line $y = x$.

Homoscedasticity: We check homoscedasticity by plotting residuals vs. fitted values. We expect a random scatter with no apparent pattern, as well as a close-to-constant variance, with the residuals centered around 0.

Based on these checks, we can select specific predictors that pass these assumption checks and/or perform transformations on variables to ensure all assumptions are affirmed.

2.2.3 Problematic Observations

To identify and deal with problematic observations, we will utilize Cook's distance to identify outliers. Given the nature of our dataset, it is expected to have outliers. For example, there could be cases of players with lower salaries from initial contracts who began improving their performance later, like rookies generally.

2.2.4 Collinearity

We will check for multicollinearity using the Variance Inflation Factor (VIF). This will help us identify inflation in the errors. We will focus on VIF values exceeding 5 to address multicollinearity. To address multicollinearity, we can re-specify the model; for example, if two or more predictors are correlated, we can remove at least one of these predictors, which should reduce the effect of multicollinearity.

2.2.6 Test Dataset

Once we have decided on our potential model and verified the necessary assumptions and conditions, we will test our model's performance. We will fit our model to the test dataset and compare the properties, such as the coefficients, with those in the training dataset. Our model will be validated if the test data fits our selected model with similar parameters and assumptions. After completing these steps we will complete all appropriate model diagnostics to affirm the goodness of our final model.

2.3 Model Violations and Diagnostics

As outlined in the Model Validation section, we will perform model violations and diagnostics during the validation of our training dataset. Additionally, after fitting our testing data and

passing model validation, we will re-evaluate these violations and diagnostics to ensure our final model meets all assumptions and conditions.

Having addressed how we perform model violations and diagnostics in the Model Validation section, we now turn to how we handle these violations.

2.3.1 Handling Violations and Diagnostics

Linearity: If non-linearity is detected, we can transform the predictors using methods such as power transformation to achieve linearity. If transformation is not feasible, we may exclude the non-linear predictor.

Normality of Residuals: If the residuals do not exhibit normality, we can attempt to transform the response variable.

Homoscedasticity: To address issues with homoscedasticity, we can apply transformations to the response variable, predictors, or both. Common transformations include square-root and natural log transformations.

Collinearity: If the Variance Inflation Factor (VIF) values are greater than 5, indicating high collinearity, we can remove at least one of the correlated predictors. If the issue persists, we can utilize dimensionality reduction techniques, such as stepwise regression, to reduce the number of parameters.

3.0 Results

This section will discuss the results achieved by the selected model.

3.1 Description of Data

Please see Appendix A for the in depth numerical summary of all our variables. After reviewing the numerical summaries we have found 3 important features:

1. A wide range of salaries (from $64,343 to $51,915,615) and positively skewed as shown in the first boxplot of Figure 1. This can lead to a violation of homoscedasticity as the larger values may act as outliers.
2. Some predictors have a high correlation between each other, for example, field goals (FG) and points (PTS) as shown in the scatter plot of Figure 1. This can lead to multicollinearity and impact the variation of the coefficients of these predictors resulting in an unstable model.
3. Performance metrics such as field goals (FG) and total rebounds (TRB) have a wide distribution as shown in the third and fourth boxplot in Figure 1. This shows each player has different roles for the team which may affect their respective salary in a nonlinear fashion.

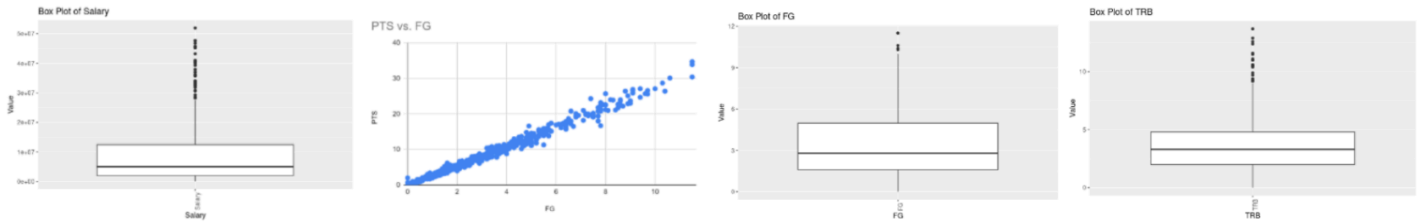

Figure 1: Plots highlighting important features of the dataset

3.2 Analysis of Results

Firstly, we went with the assumption of using the 4 best predictors that were the most correlated with salary. These were PTS, FG, FGA, FT. As seen with the statistics for Model 1 (Appendix B), most predictors are not statistically significant as p-values were above 0.2 for 3 of them. Additionally, negative predictors make the model's interpretation harder. Furthermore, the VIF are all greater than 5, which means that the data is not collinear. Based on the graphs, the model is not quite homoscedastic, and it is slightly skewed on the normality plot. From these observations, we used the stepwise method to try and better fit the model. This resulted in the adoption of the following predictors: PTS & FGA. As seen with the statistics for Model 2 (Appendix B), the model achieved a better AIC score, making it a better fit. However, based on the VIF scores being both around 37, the model indicates strong signs of collinearity so we decided to eliminate it as it violates one of our assumptions.

Secondly, we decided to take all predictors with a correlation value magnitude of > 0.5. The predictors that were selected were: GS, MP, FG, FGA, X3P, X3PA, X2P, X2PA, FT, FTA, DRB, AST, TOV, PTS. As seen in the statistics for Model 3 (Appendix B), the model has varying VIF values meaning some predictors are collinear. Thus, we decided to use stepwise analysis in order to improve the fit of the model. This provided us with the following predictors: GS, X3P, X3PA, X2P, FT, AST. As seen in the statistics for Model 4 (Appendix B) The VIF for each parameter of this model were under 5 except X3PA and X3P which were in the 40 range. From this we concluded that removing the predictors would ensure collinearity. In this case we decided to remove X3PA as it slightly improved the model's explanatory power. With this change, the VIF of each predictor for the model was under 5, meaning that it satisfies the collinear assumption.

3.3 Goodness of Model

After selecting the model with GS, X3P, X3PA, X2P, FT, AST as its predictors, we predicted the results on the validation dataset. Our model managed to achieve an R-squared value of 0.5628, which is very similar to what we observed in our training data. Thus, we can conclude that our model was not overfit to the training data, and can effectively infer real world data.

Coefficient	**Estimate**	**Pr(>|t|)**	**VIF**
Intercept	-1105087	0.117539	NA
GS	58625	0.009447	2.738302

X3P	1949099	0.001434	2.089466
X2P	1252326	0.004295	5.147592
FT	1073652	0.072636	4.641078
AST	1154240	0.000242	2.486038

Table 1: Final Model coefficient values

4.0 Discussion

The final adopted model is the stepwise linear with correlation coefficients larger than 0.5, with the 3-points shot attempts variable removed. Table 1 in the appendix describes the model's coefficients: for a hypothetical player who does not score any goals, does not distribute any assist and does not score any free throw, his salary would hypothetically be negative and equal to -1,105,087 US dollars but this intercept is not statistically significant. While such a player does not exist, it allows us to picture the contribution of each variable appropriately. The most impactful coefficient is clearly the number of 3-points shots scored: for each additional 3-point shot scored by the player, the predicted salary increases by $1,949,099. While not as large as X3P, the number of 2-points shots scored is also an excellent predictor of NBA player salary since it increases the predicted salary by $1,252,326 for each additional successful shot. Both of these results are significant at the 99% confidence level.

The model developed by (Schmidt et al.) placed points scored as one of the main predictors of salary as it is associated with a 7.7% increase in salary. This is quite coherent with the results we obtain here with more recent data as predictors of shots scored, both at 2 or 3-points range, are statistically significantly associated with salary.

These results must however be read with caution as predictors of points scored are correlated with other variables: the correlation coefficient for attempted shots at both the 2 and 3-point range is 0.987. (Lyons Jr. et al.) found the percentage of scored shots to be statistically significant in their model. While the percentage of 3-points scored (relative to total attempts) was considered at some point of the analysis, it was taken out of the model but it is still moderately correlated (coefficient is 0.443) with 3-points scored.

Another limitation of the model is its relatively low explanatory power: variables only explain 57% of the NBA salary variations and

A final important limitation is the issue of homoscedasticity. The plot of fitted values vs. residuals (Figure 6) reveals that the first half of the data does not exhibit constant variance. Given the complex correlation within our data, a simple transformation does not provide a significant improvement. More sophisticated solutions, such as GLS, could complicate the interpretation of the results increasing in the complexity of the model. However, addressing this limitation is crucial, as heteroscedasticity can lead to inefficient estimates and invalidate statistical tests.

5.0 References

Jackson Jr., Newton, et al. "Determinants of NBA Player Salaries." The Sport Journal, 13 July 2018, thesportjournal.org/article/determinants-of-nba-player-salaries/.

NBA. "2023-24 NBA Player Stats: Per Game." Basketball Reference, 2024, www.basketball-reference.com/leagues/NBA_2024.html.

Schmidt, Martin, and David Berri. "(PDF) Does One Simply Need to Score to Score?" ResearchGate, 2007, www.researchgate.net/publication/5142758_Does_One_Simply_Need_to_Score_to_Score .

6.0 Appendix

Appendix A: Numerical Summary of each Variable

	Salary	Age	Games Played	Games Started	Minutes Played	Field Goals	Field Goals Attempted	Field Goal Percentage	Three Point Field Goals
Minimum	64,343	19	2	0	2.3	0	0.3	0	0
1st Quartile	2,031,426	23	38	1	13	1.6	3.6	42.13	0.3
Median	5,013,400	25	57	12	20	2.8	6.15	45.45	0.8
Mean	9,946,186	26.07	52.7	25.38	20.58	3.508	7.452	46.17	1.055
3rd Quartile	12,476,250	29	71	49.75	28.48	4.975	10.1	50	1.6
Maximum	51,915,615	39	84	82	37.8	11.5	23.6	74.7	4.8

Table 2: Summary of First 9 Variable

	Three Point Field Goal Attempts	Three Point Field Goal Percentage	Two Point Field Goals	Two Point Field Goal Attempts	Two Point Field Goal Percentage	Effective Field Goal Percentage	Free Throws	Free Throws Attempted	Free Throw Percentage
Minimum	0	0	0	0	0	0	0	0	0
1st Quartile	1.125	29.55	1	1.8	49.2	49.8	0.5	0.7	70
Median	2.5	35	1.8	3.3	53.8	53.6	0.9	1.3	77.8
Mean	2.936	32.36	2.451	4.517	53.31	53.08	1.413	1.809	75.48
3rd Quartile	4.4	38.5	3.475	6.2	58.1	57.7	1.8	2.4	83.37
Maximum	11.8	100	11	18.3	100	74.7	10.2	11.6	100

Table 3: Summary of Next 9 Variables

	Offensive Rebounds	Defensive Rebounds	Total Rebounds	Assists	Steals	Blocks	Turnovers	Personal Fouls	Points
Minimum	0	0	0	0	0	0	0	0	0
1st Quartile	0.4	1.5	2	0.9	0.4	0.2	0.5	1.1	4.4
Median	0.7	2.5	3.3	1.5	0.6	0.3	0.9	1.6	7.3
Mean	0.9363	2.809	3.737	2.223	0.651	0.445	1.098	1.63	9.481
3rd Quartile	1.2	3.6	4.8	3	0.9	0.6	1.5	2.1	13.15
Maximum	4.6	10.1	13.7	10.9	2.1	3.6	4.4	3.6	34.7

Table 4: Summary of Last 9 Variables

Appendix B: Models

Model 1 - Top4 Predictors:

Multiple R-squared	Adjusted R-squared	F-statistic	P-value	AIC
0.54	0.5347	103.3 on 4 and 352 DF	< 2.2e-16	12294.7

Coefficient	Estimate	Pr(>|t|)	VIF
Intercept	-900837	0.21199	NA
PTS	2627688	0.00186	213.33444
FG	-910500	0.50757	75.11990
FGA	-1301991	0.02221	53.24110
FT	-1039858	0.26323	10.75597

Table 5: Model 1 values

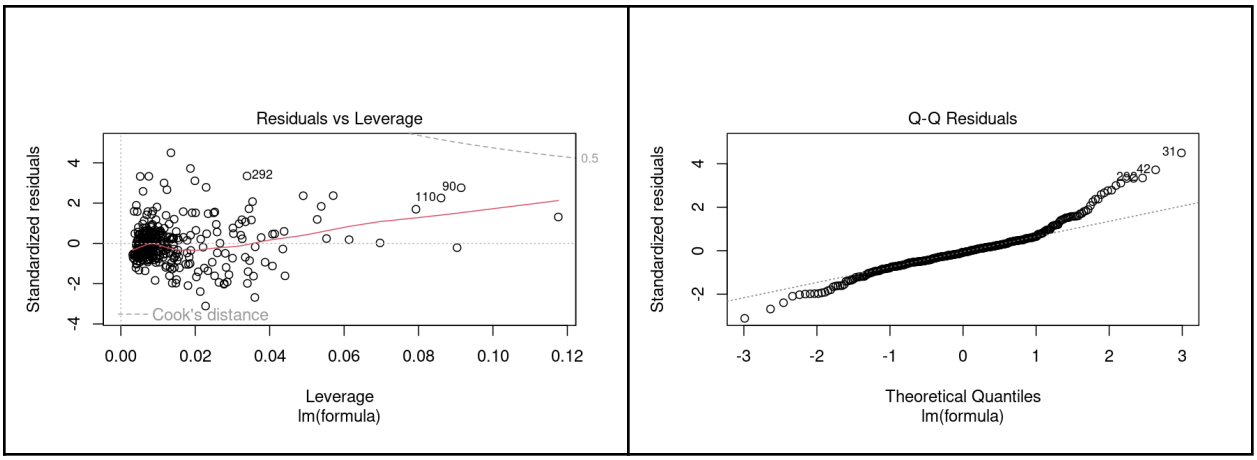

Figure 2: Model 1 residual plots

Model 2 - Stepwise from the top4 predictors:

Multiple R-squared	**Adjusted R-squared**	**F-statistic**	**P-value**	**AIC**
0.5383	0.5357	206.4 on 2 and 354 DF	< 2.2e-16	12291.98

Coefficient	**Estimate**	**Pr(>|t|)**	**VIF**
Intercept	-822178	0.2452	NA
PTS	1862772	2.34e-07	37.96664
FGA	-965544	0.0442	37.96664

Table 6: Model 2 values

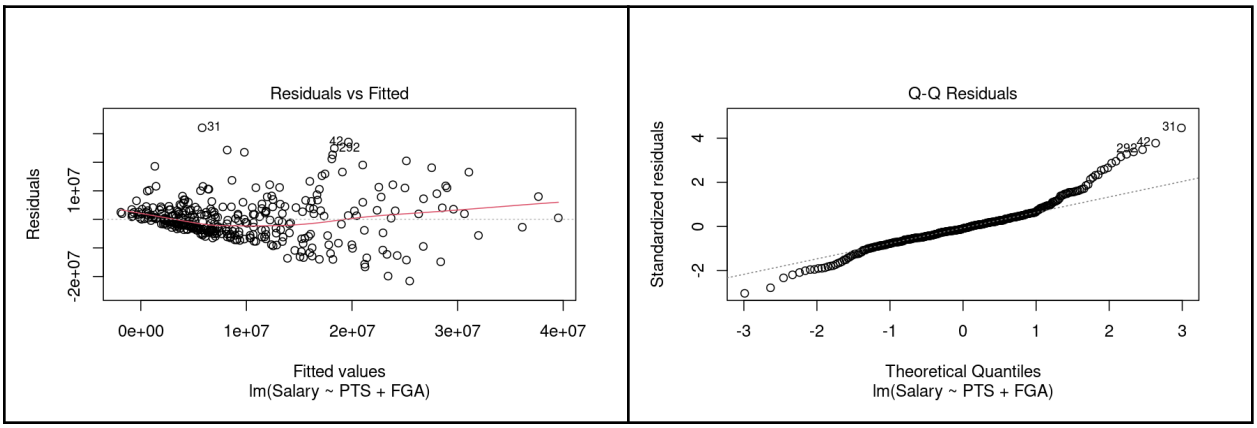

Figure 3: Model 2 residual plots

Model 3 - all significant parameters used (cor(salary) > 0.5):

Multiple R-squared	**Adjusted R-squared**	**F-statistic**	**P-value**	**AIC**
0.578	0.5607	33.46 on 14 and 342 DF	< 2.2e-16	12283.9

Coefficient	**Estimate**	**Pr(>|t|)**
Intercept	117805	0.921917
GS	52809	0.041879
MP	-55736	0.646678
FG	-2864124	0.796086
FGA	4582696	0.550402
X3P	9068532	0.374805
X3PA	-6929513	0.375007
X2P	4655889	0.525711
X2PA	-5520918	0.473113
FT	3244399	0.558161
FTA	-1989170	0.273198
DRB	275930	0.507200
AST	1561052	0.000149
TOV	-1344777	0.298267
PTS	609132	0.905821

Table 7: Model 3 values

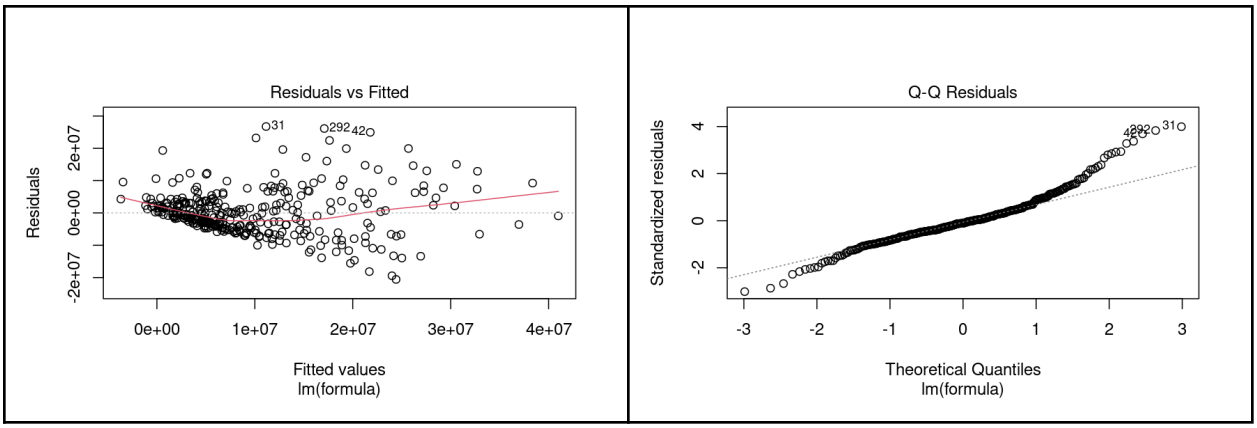

Figure 4: Model 3 residual plots

Model 4 - Stepwise model from (cor(salary) > 0.5):

Multiple R-squared	**Adjusted R-squared**	**F-statistic**	**P-value**	**AIC**
0.5713	0.564	77.74 on 6 and 350 DF	< 2.2e-16	12273.5

Coefficient	**Estimate**	**Pr(>|t|)**	**VIF**
Intercept	-335378	0.653197	NA
GS	53612	0.016713	2.754843
X3P	9522139	0.000429	41.598820
X3PA	-3061030	0.003951	42.079311
X2P	1236579	0.004383	5.148407
FT	1317094	0.027768	4.734913
AST	1187742	0.000138	2.489533

Table 8: Model 4 coefficient values

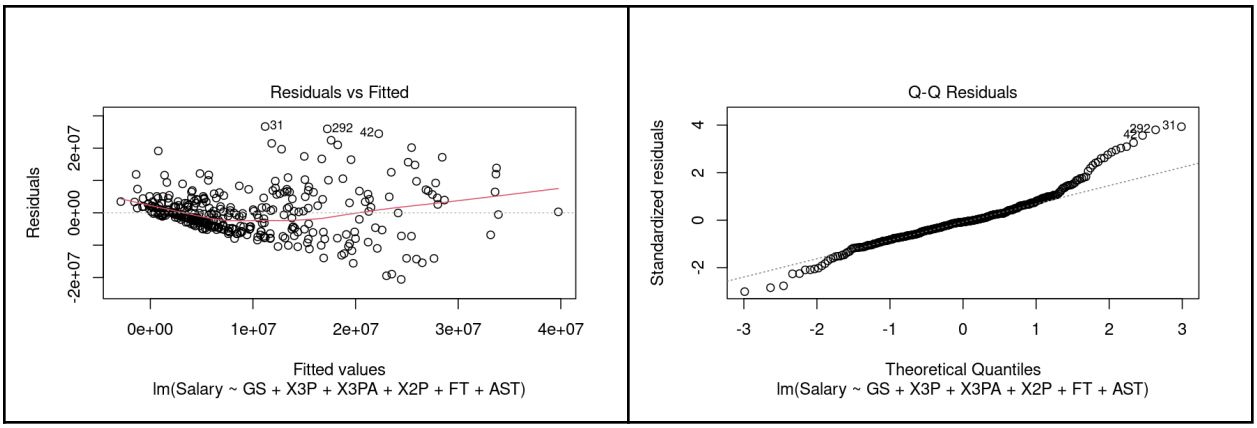

Figure 5: Model 4 residual plots

Final Model - Stepwise with cor > 0.5, X3PA removed:

Multiple R-squared	**Adjusted R-squared**	**F-statistic**	**P-value**	**AIC**
0.561	0.5548	89.71 on 5 and 351 DF	< 2.2e-16	12279.98

Table 9: Final model values

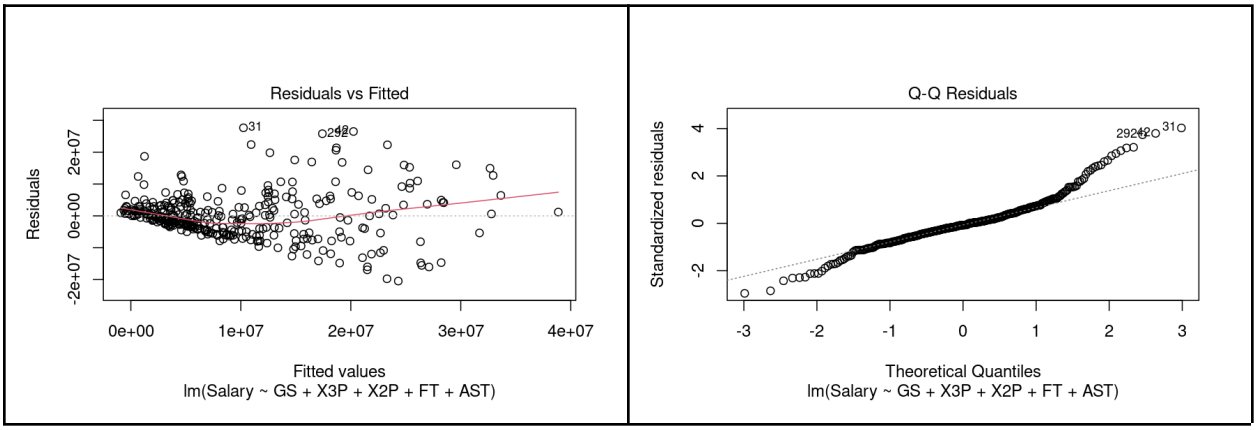

Figure 6: Model 5 residual plots

Predicted R^2 = 0.562807992129736