Final Project Part 3

STA302H1: Methods of Data Analysis I

Christoffer Tan (1008740445), Janis Joplin (1009715051), Razan Ahsan Rifandi (1009562108)

Contributions

We worked on all parts together, but each section was assigned to a specific team member as the person in charge of that section.

- **Christoffer**: Introduction, Conclusion and Limitations
- **Janis**: Results
- **Razan**: Ethics Discussion, Methods

Introduction

NBA player salaries reflect athletic performance, career achievements, and roles, making them a key focus for sports economists. This study examines: **To what extent can player performance metrics and achievements predict NBA salaries using linear regression?**

Previous research highlights the strong link between performance metrics and salary. Rosen et al. (2013) identified field goal percentage and points per game as significant predictors, along with assists and rebounds (adjusted R-squared: 0.613). These findings emphasize the value of well-rounded performance metrics and career milestones in explaining salary variations. Sigler and Sackley (2000) found that offensive and defensive attributes significantly influence salaries, underscoring the value of well-rounded performance. Bodvarsson and Brastow (1998) highlighted All-Star participation and positional roles, especially for centers and forwards, as key salary predictors. These findings collectively suggest that performance metrics, career achievements, and player roles are strongly associated with salary levels.

Building on these findings, our study incorporates metrics such as points, assists, rebounds, minutes played, efficiency measures, positions, and All-Star participation to develop a predictive linear model for NBA salaries. Linear regression is well-suited for this analysis, as it quantifies relationships between predictors and a dependent variable, offering interpretable coefficients and statistical measures of model performance. By integrating these factors, this study aims to identify key predictors and provide actionable insights into NBA salary structures.

Methods

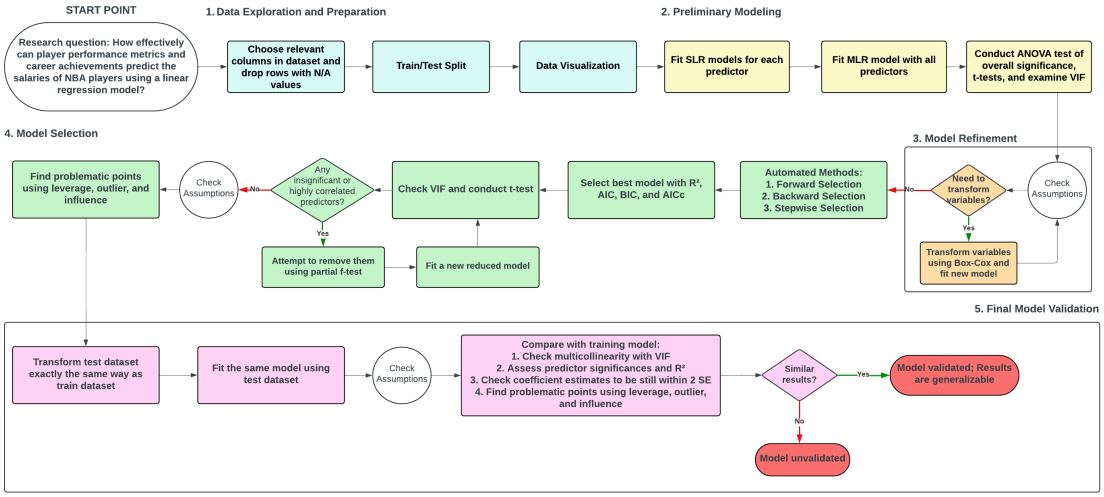

Figure 1: Flowchart describing the methods of analysis of the research

Data Exploration and Preparation

We selected relevant predictor variables for our research question. We then searched for missing values in the dataset and removed observations containing them. The dataset was then randomly split into training and testing data (50/50) for validation. EDA involved visualizing numerical variables with histograms and boxplots and categorical variables with barplots to analyze data distribution. Outliers in the dataset were identified by finding observations that falls outside 1.5 IQR.

Preliminary Modeling

We fit SLR models for each predictor variable to assess each individual variable's relationship with salary and determine statistical significance using t-test and R^2 value. We then fitted a MLR model with all predictors and conduct ANOVA test of overall significance and t-test to evaluate statistical significance of each predictor within the full model. We also assess multicollinearity in the full model through VIF values.

Model Refinement

We checked the additional conditions of MLR: conditional mean response and conditional mean predictors using a scatterplot of response vs fitted (checked with random diagonal scatter or non-linear trend) and pairwise scatterplot of predictors (check with lack of curves or non-linear patterns). Violations would indicate that the residual plots will be unreliable. We then verified linear regression assumptions: linearity, uncorrelated errors, constant variance, and normality. The first three are identified with systematic patterns, clusters, and fanning in the residual plots, while the latter is identified by stark deviations in the normal QQ line. Violations would indicate that our model was inappropriate and some variables might need to be transformed. Finally, we transformed the variables with Box-Cox and simple `powerTransform` if deemed necessary and reverified the assumptions with the transformed model.

Model Selection

We built models by employing three automated procedures using the `stepAIC` function from the R MASS library which uses AIC as measure: forward, backward, and stepwise selection.

We compared the three models using adjusted R-squared, AIC, BIC, and AICc to pick the best model. To improve adjusted R-squared, we evaluated VIF values to check for multicollinearity and attempt to remove highly correlated predictors by conducting partial F-test and remove them if they are not statistically significant. This is repeated until there are no significant multicollinearity in the model. Lastly, we reverified MLR assumptions to determine whether our model is appropriate and identified problematic points using leverage, outlier, and influence that might affect our model.

Final Model

We started by transforming the test data with the same transformation we applied to training data and create test model with testing data. MLR assumptions were then reverified in the test model. We examined VIF values of test model and compared it to train model. Then, we compared significant predictors and evaluated adjusted R-squared values between training and test models. We verified that coefficient estimates remained within two standard errors of their training model values and examined problematic points in the test dataset. Finally, we validated the model and determine its generalizability using the results.

Results

Data Exploration and Preparation

After removing 66 observations with missing values in `eFG.` or `Salary`, 1342 observations remained. The dataset was split equally into training and test datasets of 671 observations each (50% training set and 50% test set). Visualizing the train dataset revealed that most numerical predictors and the response variable (`PTS`, `AST`, `BLK`, `TRB`, `Salary`) were right-skewed with outliers, while `eFG.` was symmetric but had outliers on both sides. `MP` and `STL` had normal distributions with minimal outliers. Identifying outliers using $1.5 \times \text{IQR}$, the counts were: `AST` (47 outliers), `eFG.` (43 outliers), `BLK` (40 outliers), `Salary` (34 outliers), `TRB` (31 outliers), `PTS` (18 outliers), `STL` (7 outliers), and `MP` (none). Among categorical predictors, `Pos1` was uniformly distributed across positions (C, PF, PG, SF, SG), and only 37 players participated in the All-Star match (`Play = Yes`), compared to 634 who did not (`Play = No`).

Preliminary Modeling

Simple linear regression for each predictor showed all predictors were statistically significant, with `PTS` and `AST` having relatively larger R^2 values. A multiple linear regression (MLR) model including all predictors yielded a significant ANOVA p-value, confirming a linear relationship for at least one predictor. T-tests indicated that all predictors except `eFG.`, `MP`, and `STL` were significant in the presence of other predictors, and multicollinearity was severe for some predictors.

Model Refinement

In Figure 3, the random scatter in the Salary vs Fitted plot and the absence of non-linear patterns in pairwise predictor scatterplots suggest residuals can be used to assess linear regression assumptions.

- Linearity: no systematic patterns or function of predictors, **not violated**
- Uncorrelated errors: no clustering or sequential patterns, **not violated**
- Constant variance: significant fanning pattern, **violated**

- Normality: stark deviation from the diagonal line, **violated**

To address violations, numerical predictors with zero or negative values were shifted by +1, followed by Box-Cox transformations using `powerTransform` with these λ values:

1. `PTS`: lambda = 0 $\rightarrow \log(\texttt{PTS})$
2. `AST`: lambda = −0.5 $\rightarrow \frac{1}{\sqrt{\texttt{AST}}}$
3. `BLK`: lambda = −2 $\rightarrow \frac{1}{\texttt{BLK}^2}$
4. `TRB`: lambda = 0 $\rightarrow \log(\texttt{TRB})$
5. `eFG.`: lambda = 4 $\rightarrow (\texttt{eFG.})^4$
6. `MP`: lambda = 0.75 $\rightarrow \texttt{MP}^{0.75}$
7. `STL`: lambda = −1 $\rightarrow \frac{1}{\texttt{STL}}$
8. `Salary`: lambda = 0.25 $\rightarrow \sqrt[4]{\texttt{Salary}}$

We used these lambda values directly, as they are optimized to maximize the likelihood of the transformed variables satisfying linear model assumptions. Next, we reassess the MLR and linear regression assumptions. Plots after transformation (Figure 4) confirmed that MLR assumptions, linearity, and uncorrelated errors remain satisfied. Furthermore, the absence of fanning and minimal QQ plot deviations indicate that constant variance and normality assumptions are now met.

Model Selection

Automated selection methods yielded these models:

1. Forward selection: `PTS`, `AST`, `BLK`, `eFG.`, `MP`, `Pos1`, `Play`
2. Backward/Stepwise selection: `PTS`, `AST`, `TRB`, `eFG.`, `MP`, `Pos1`, `Play`

The second model was selected due to its slightly better metrics: adjusted R^2 (0.462), AIC (3139.334), BIC (3188.93), and AICc (3139.806), compared to the first model's adjusted R^2 (0.4616), AIC (3139.917), BIC (3189.513), and AICc (3140.389). T-tests show that all predictors, except `TRB`, are statistically significant.

After detecting severe multicollinearity in the model due to `MP` (VIF = 11.59), this predictor was excluded. However, multicollinearity persisted, particularly for `PTS` (VIF = 5.19), followed by `AST` (4.61) and `TRB` (4.55). Since `PTS` is widely acknowledged in the literature as a key predictor of salary, we employed a partial F-test instead to evaluate the necessity of `TRB`, as it was the only non-statistically significant predictor. Comparing a model excluding `TRB` to the previous model, the ANOVA test results were not statistically significant, indicating no significant linear relationship between `Salary` and `TRB`. Consequently, `TRB` was removed, leaving the final model with `PTS`, `AST`, `eFG.`, `Pos1`, and `Play`.

Final Model

Plots for the final model (Figure 5) confirmed the validity of linear regression assumptions. The Salary vs Fitted plot displayed a random scatter, and no nonlinear patterns were observed in the pairwise scatterplots of predictor. Residual plots and QQ plot further revealed:

- Linearity: no systematic patterns or function of predictors, **not violated**
- Uncorrelated errors: no clustering or sequential patterns, **not violated**
- Constant variance: no fanning pattern, **not violated**

- Normality: minor deviation from the diagonal line, aside from one problematic point, **not violated**

The final model resolved severe multicollinearity, with all predictors having VIF values below 5. Diagnosing problematic points identified 53 leverage points and 24 outliers. The model had no influential point on all fitted values but did have 24 influential points on their own fitted values. Influential points for coefficient estimates were:

- β_0: 40
- β_1: 52
- β_2: 47
- β_3: 34
- β_4: 70
- β_5: 48
- β_6: 55
- β_7: 53
- β_8: 14

Conclusion and Limitations

The final model uses Box-Cox transformed variables to predict NBA player salaries is:

$$\sqrt[4]{\texttt{Salary}} = \beta_0 + \beta_1 \log(\texttt{PTS} + 1) + \beta_2 \left(\frac{1}{\sqrt{\texttt{AST} + 1}} \right) + \beta_3 \left((\texttt{eFG.} + 1)^4 \right) + \beta_4 I_{\texttt{Pos1} = \texttt{PF}} + \beta_5 I_{\texttt{Pos1} = \texttt{PG}} + \beta_6 I_{\texttt{Pos1} = \texttt{SF}} + \beta_7 I_{\texttt{Pos1} = \texttt{SG}} + \beta_8 I_{\texttt{PlayYes}}$$

Coefficients	Estimate	Std. Error	T-value	P-value
Intercept	46.7853	5.8911	7.942	< 0.001
$\log(\texttt{PTS} + 1)$	9.8242	1.2061	8.146	< 0.001
$\left(\frac{1}{\sqrt{\texttt{AST} + 1}} \right)$	-19.4321	5.3472	-3.634	< 0.001
$(\texttt{eFG.} + 1)^4$	-0.9147	0.428	-2.137	0.033
Pos1PF	-5.3281	1.282	8.146	< 0.01
Pos1PG	-8.918	1.627	-5.481	< 0.001
Pos1SF	-3.9817	1.3801	-2.885	0.004
Pos1SG	-8.0479	1.2697	-6.339	< 0.001
PlayYes	6.2993	1.94	3.247	0.0012

Table 1: Final model summary on train data

This predictive model effectively answers the research question by demonstrating that NBA player salaries are significantly influenced by performance metrics (PTS, AST, and eFG%), positional roles, and career achievements (All-Star participation) with an adjusted R^2 value of 0.45. Based on Table 1, scoring ability (PTS) emerges as the most significant predictor, while assists (AST) shows an inverse relationship with salaries, highlighting a potential trade-off between scoring and playmaking. Efficiency field goal percentage (eFG%) has a smaller negative effect on salaries. Positional roles reveal that centers typically earn more compared to power forwards, shooting guards, point guards, and small forwards, aligning with their physical demands and strategic value. All-Star participation is positively significant, showing its importance as a career milestone in determining compensation. This model is important as it provides valuable insights into the key factors driving NBA player salaries, aiding in more informed contract negotiations, team budgeting, and player evaluation.

These results are consistent with existing literature:

- Rosen et al. (2013): similarly demonstrates that points per game and assists are significant predictors
- Sigler and Sackley (2000): confirms that offensive metrics (PTS, AST, eFG%) significantly influence salaries
- Bodvarsson and Brastow (1998): similar to our findings on All-Star participation and positional roles as key salary predictors

Coefficients	Estimate	Std. Error	T-value	P-value
Intercept	59.8541	5.6526	10.589	< 0.001
$\log(\texttt{PTS} + 1)$	7.9015	1.1337	6.969	< 0.001
$\left(\frac{1}{\sqrt{\texttt{AST} + 1}}\right)$	-34.7525	5.1694	-6.723	< 0.001
$(\texttt{eFG.} + 1)^4$	-0.6062	0.2537	-2.389	0.017154
Pos1PF	-2.4083	1.3104	-1.838	0.066524
Pos1PG	-14.6350	1.5818	-9.252	< 0.001
Pos1SF	-6.3406	1.4185	-4.470	< 0.001
Pos1SG	-7.5369	1.3622	-5.533	< 0.001
PlayYes	7.7410	2.0375	3.799	0.000158

Table 2: Final model summary on test data

Validation on the test dataset produced similar results, with slightly lower VIF values compared to the training dataset, with AST (3.84) and PTS (3.17) have the highest VIF values. Most predictors remain statistically significant, except for the Power Forward (PF) position.

Coefficient estimates stay within acceptable ranges, with differences generally within three standard errors. The adjusted R-squared increased slightly to 0.48, confirming the model's robustness. As shown in Figure 2, diagnostic tests presents no major violations of assumptions. Problematic data points are similar in both datasets, with 47 leverage points and 23 outliers. There were no influential points on all fitted values, but 16 influential points were detected for their own fitted values. Specific influential points were still identified for each coefficient estimate.

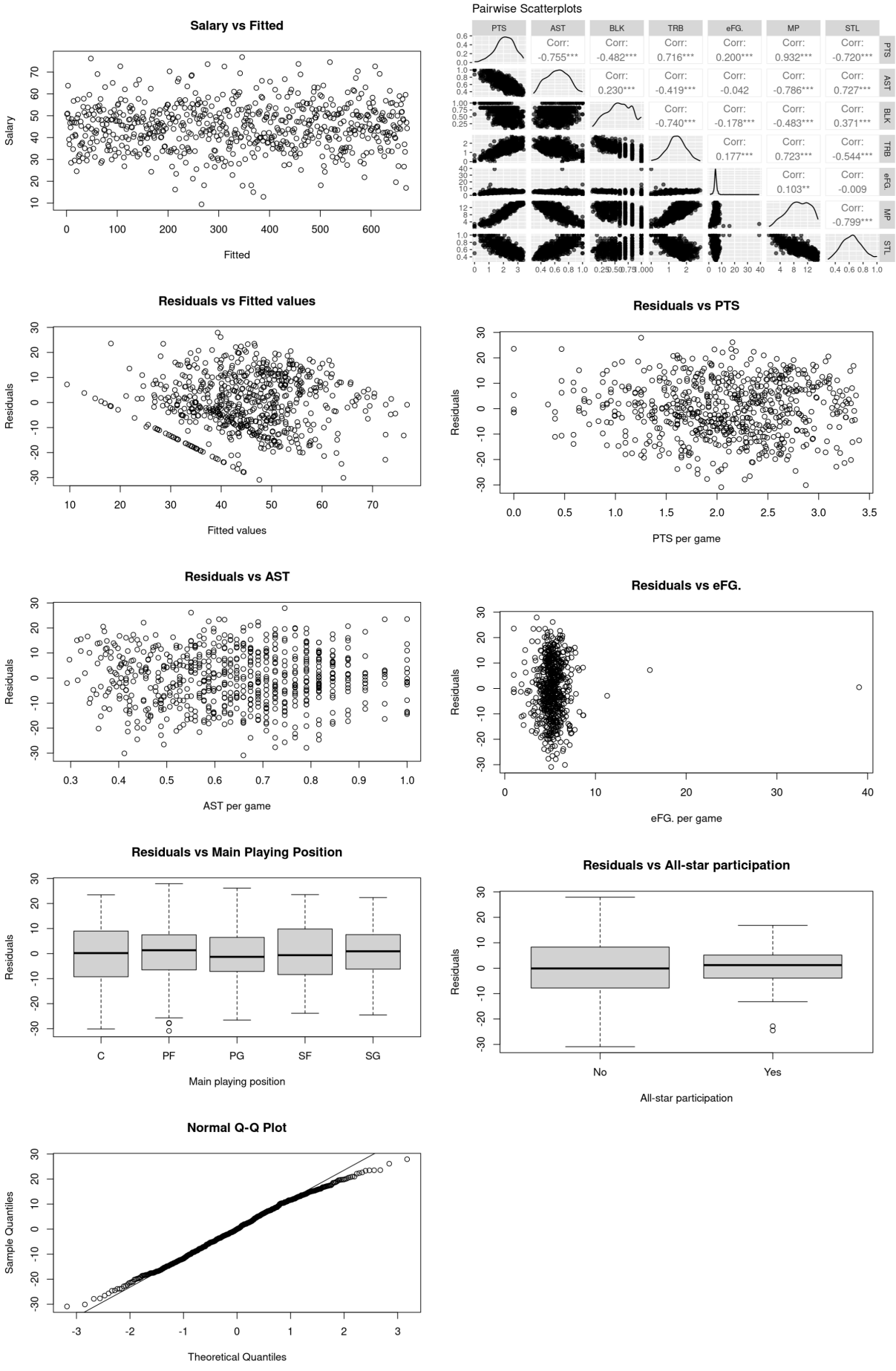

Figure 2: Plots to check assumptions on final test model

Limitations include potential issues with rookie salary caps, inconsistent player performances, and the need for data transformation for Box-Cox compatibility. Additionally, systemic biases like racial discrimination might affect salary predictions. Future research could address these challenges by exploring nonlinear models or incorporating more contextual factors, such as team dynamics. Despite these limitations, the model offers valuable insights into salary determinants, emphasizing the role of performance and career milestones.

Ethics Discussion

Methodological choices in statistical analysis have significant ethical implications, especially in model selection procedures. As our analysis considers ten predictor variables for the multiple linear regression model to answer our research question, the complexity of variable selection creates a significant risk for human bias and error. Automated selection methods represent a more ethically sound approach because they provide objective criteria for model building, reducing the possibility of personal bias influencing results. Therefore, we considered automated selection methods to be preferable to manual selection for our analysis.

Our analysis used three automated selection methods: forward, backward, and stepwise selection. These automated selection methods has several ethical benefits over manual selection methods. First, they prevent personal biases, which could arise from manually entering variables into the model based on subjective beliefs. Second, they handle the complexity of evaluating variable combinations of many predictor variables in an efficient and systematic manner. This prevents human oversight that could occur from using manual selection methods. Third, the process is transparent and documentable, which maintains the scientific integrity of our project. We have properly documented the code we used for our research and utilized a widely-used R library MASS for the automated selection methods. This ensured that our results were reproducible. Lastly, the use of automated methods also prevents potentially harmful statistical practices such as p-hacking and data manipulation in order to achieve desired results.

Bibliography

- Lyons Jr., R., Jackson Jr., E. N., & Livingston, A. (2015). *Determinants of NBA player salaries*. The Sport Journal. https://doi.org/10.17682/sportjournal/2015.019.
- Sigler, K. J., & Sackley, W. H. (2000). *NBA players: Are they paid for performance?*. Managerial Finance, 26(7), 46–51. https://doi.org/10.1108/03074350010766783.
- Bodvarsson, O. B., & Brawstow, R. T. (1998). *Do employers pay for consistent performance?: evidence from the NBA*. Economic Inquiry, 36(1). https://go-gale-com.myaccess.library.utoronto.ca/ps/i.do?p=BIC&u=utoronto_main&id=GALE%7CA20611177&v=2.1&it=r&sid=summon
- Ratto, Davide. (2019). *NBA Players 2016-2019*. [Data set] https://www.kaggle.com/datasets/davra98/nba-players-20162019

Appendix A

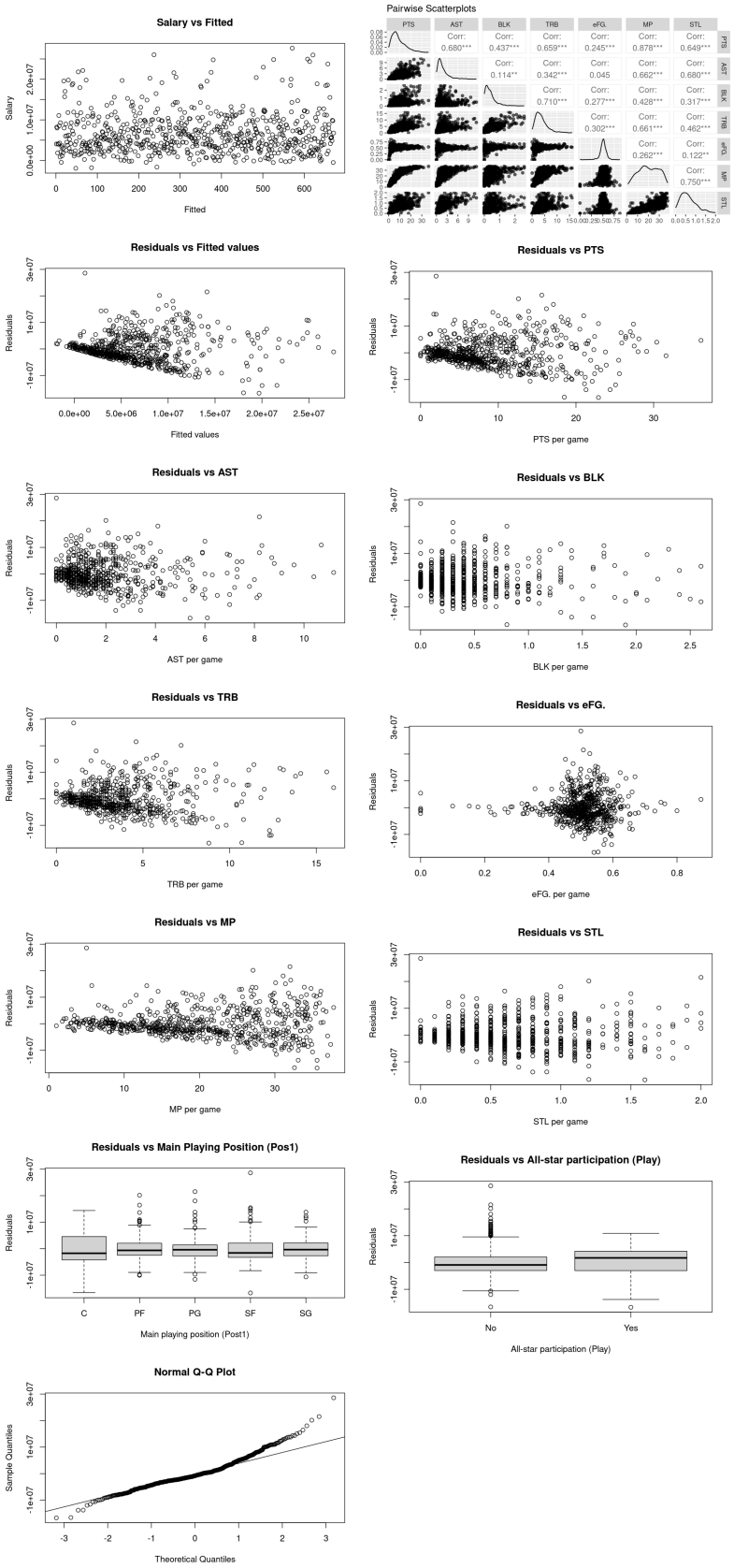

Figure 3: Plots to check assumptions before any transformation

Appendix B

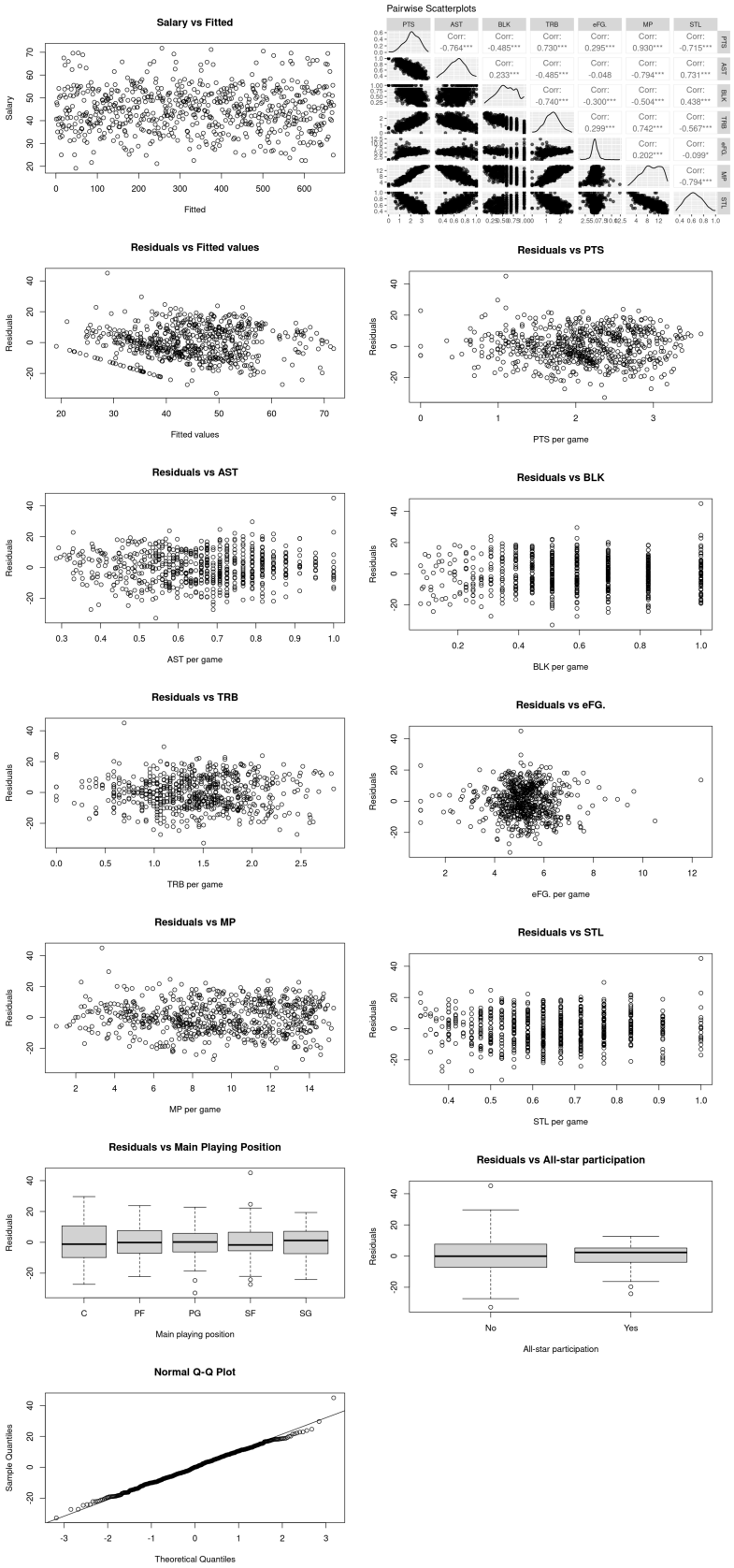

Figure 4: Plots to check assumptions after box-cox transformation

Appendix C

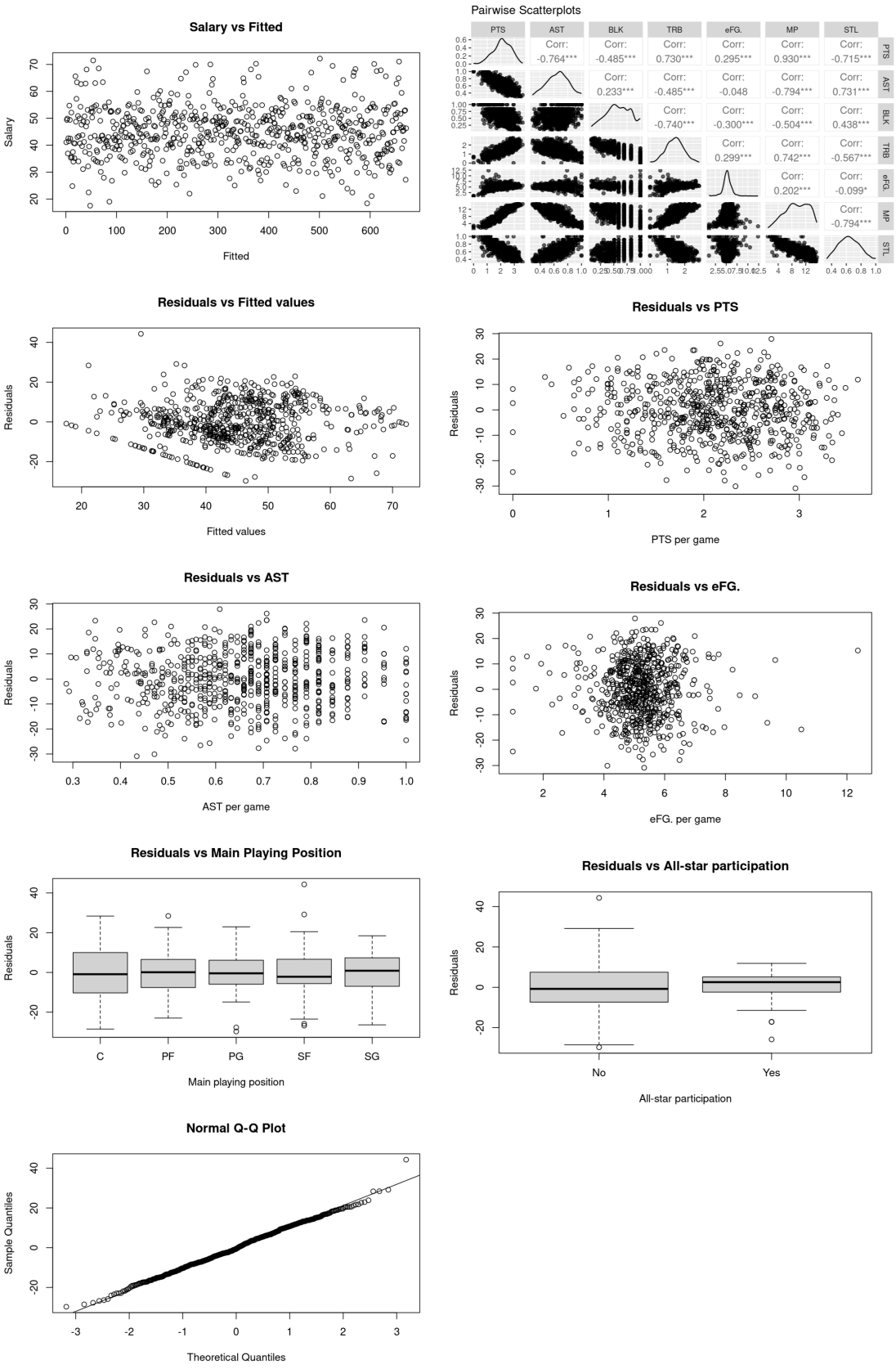

Figure 5: Plots to check assumptions on final train model