1. Introduction

During 2022/2023 season, 30 teams in NBA have a total revenue amounting to 10.58 billion U.S. dollars[1]. Understanding the relationship of performance statistics and the winning probability can help NBA teams to make better game strategies to maximize the winning probability, which can further increase the profits of a team. Hence, our purpose is to fit a descriptive model to maximize the interpretability of the relationship between statistics and the outcome.

Previous study ensure the feasibility and effectiveness of using performance statistics such as free throw percentage, 3-points percentage, rebounds, assists, to predict the outcome of a NBA game. Their choice of predictors overlap most of our interested statistics in the data set. However, previous study use Naive Bayes, bi-variate normal mean regression model[2][3][4], while we use logistic model to predict the winning probability.

Therefore, the research question is 'How the performance statistics of home team can affect the its wining probability in a NBA game. The logistic model with logit link is chosen due to its ability of predicting the winning probability of binary response. Moreover, logistic model can help to identify statistically significant predictors, and the change of the odds ratio when a specific predictor change and other predictors stay unchanged can be quantified.

2. Method

Data Cleaning and EDA

Step one, the dataset is cleaned by removing N/A values. All the exposures and confounders are identified by reasoning and literature review. Exploratory data analysis(EDA) of these predictors and response is conducted to show the important characteristics of data and identify potential problems that may impact the generalizability to make reliable inference.

Model Assumptions checking

A primary model is fitted with all predictors. Since our aim is to fit a logistic regression model, three assumptions are required to be checked to ensure the liability of inference. First assumption is to ensure the response follows binary distribution. Second assumption is to ensure the logit link is used so that the linearity is met. Third assumption of independence is to ensure observations are independent of each other.

Model selection

When redundant predictors are included in the model, the prediction variance is increased which leads to worse accuracy. Hence, the methods of measuring goodness of fit are used to find the optimal model. First two metrics are AIC(Akaike information criterion) and BIC(Bayesian information criterion). A optimal model is

found by searching the smallest AIC or BIC by fitting models of different number of predictors in forward, backward, or stepwise direction. Third approach is to use LASSO(least absolute shrinkage and selection operator) to find a optimal model, achieved by applying penalty on large coefficient estimates to shrink them to zero. As our goal is to maximize the interpretability, we will prefer a model with less predictors while making decision.

If a selected small model is a subset of another selected large model, then likelihood ratio test can be used to make choice by performing hypothesis test and checking the p-value.

Model diagnostics

After finding the optimal model, a complete diagnostics is conducted to identify underlying problem, which helps to assess the reliability and limitation of the model. Firstly, influential observations on the estimated coefficients can be found by checking if a threshold is exceeded by dfbetas, a measurement of the effect of each observation on the estimated coefficients. Then further investigation on if these points skew the model prediction is conducted to decide whether to transform the predictors or remove these observations. Secondly, the presence of any patterns, extreme dispersion, or large deviation from horizontal axis in deviance residual plots indicates inadequate goodness of fit and potential assumptions violation. This can be addressed by transformation. Thirdly, the severeness of multicollinearity is checked by calculating the VIF (variance inflation factor) of each predictor. If VIF is greater than 5, then multicollinearity exists, which can addressed by transformation or removal of predictors.

Model validation

Evaluating the prediction accuracy and detecting the overfitting can be accomplished by resampling method, cross validation, which repeatedly partitions the original dataset into a training set to train the model, and a test set to evaluate the accuracy. Then, the performance of the model is evaluated by checking how closely the bias corrected line aligns with the diagonal line in calibration plot. The diagonal line represents the predicted probability and observed probability are same.

Also, the performance can be assessed by calculating the area(AUC) under receiver operating characteristic (ROC) curve, which represent the chance of the model can discriminate the binary outcome correctly.

3. Result

Data Cleaning and EDA:

There are 25754 observations and 28 variable initially. The dataset is cleaned by removing missing values, and 25662 observations are left. Then binary variable ‘if home team wins’ is chosen as the response. Six predictors including exposure and

confounders are chosen based on our interest, reasoning, or the information in literature. For example, potential confounder, 3-points percentage of home team have effect both on the outcome and the scored points, because high 3-points percentage means higher points, which increases the winning probability. Corresponding histograms and boxplots can be found at [Appendix A].

Table 1: Summary table of characteristics of 6 predictors and response.

Predictor	**Min**	**1st quartile**	**Median**	**Mean**	**3rd quartile**	**max**	**Std.Error**
1. Points of home team **(Exposure)**	36.0	94.0	103.0	103.5	112.0	168.0	13.2834
2. Field Goal percentage of home team **(Exposure)**	0.2500	0.4220	0.4600	0.4607	0.5000	0.6840	0.0567
3. Free throw percentage of home team **(Confounder)**	0.1430	0.6970	0.7650	0.7604	0.8330	1.0000	0.1007
4. 3-points percentage of home team **(Confounder)**	0.000	0.286	0.357	0.356	0.429	1.000	0.1112
5. Number of assists of home team **(Exposure)**	6.00	19.00	23.00	22.82	26.00	50.00	5.1933
6. Number of rebounds of home team **(Exposure)**	15.00	39.00	43.00	43.37	48.00	72.00	6.6258

If home team wins (Response)	**Counts**
0 (home team lose)	10907
1 (home team win)	15645

Model Assumption

Firstly, from Table 1, it is clear the response is binary. Secondly, as logistic model with logit link is used to predict the probability, so the linearity is satisfied. Plots in [Appendix B] also prove the linearity is verified. Thirdly, observations, the result and the statistics of every game, are independent from each other because every game is a separate event with its own set of conditions. Hence, independence assumption hold. All assumptions are satisfied, so reliable inference are guaranteed.

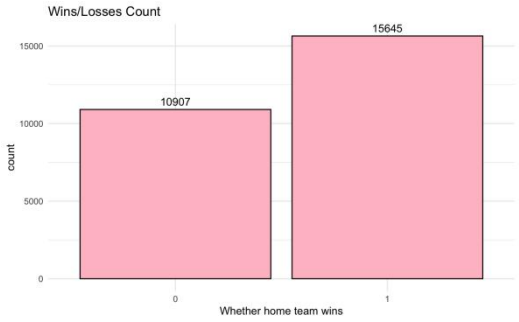


Figure 1: the distribution of the binary response

Model Selection

Firstly, a primary model with all predictors is fitted, and '#assists of home team' is found out insignificant from Table 2. This indicates it may be potentially insignificant. Then the model selected by 'backwards AIC' is the same as the primary model. However, the model selected by 'Backwards BIC' removes '#assists of home team', because BIC method penalizes more on the number of predictors than AIC when selecting. Lastly, predictors selected by LASSO are the same as the primary model. As the BIC_model is a subset of the primary model, a likelihood test is conducted. P-value(0.685) is bigger than the threshold=5%, which means the null hypothesis ($\beta_{\#assist}=0$) is not rejected, so '#assists of home team' is insignificant. Additionally, based on the principle of maximizing the interpretability, the BIC_model is chosen as the final model since it contains less predictors.

Table 2. summary of all models in the process of model selections

1. **Primary model:**

$$\log\left(\frac{E(Y|\mathbf{X})}{1-E(Y|\mathbf{X})}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6$$

	Coefficient estimate	Std.Error	z value	p-value	Significant?	VIF
Intercept	-21.389 (β_0)	0.300	-71.279	$2*10^{-16}$	Yes	
Points of home team(X1)	-0.014 (β_1)	0.002	-7.167	$7.69*10^{-13}$	Yes	1.979
Field Goal percentage of home team(X2)	26.037 (β_2)	0.494	52.765	$2*10^{-16}$	Yes	2.211
Free throw percentage of home team(X3)	3.397 (β_3)	0.164	20.710	$2*10^{-16}$	Yes	1.126
3-points percentage of home team (X4)	3.940 (β_4)	0.165	23.941	$2*10^{-16}$	Yes	1.181
#Assists of home team(X5)	0.002 (β_5)	0.004	0.405	0.685	No	1.441
#Rebounds of home team (X6)	0.169 (β_6)	0.003	55.033	$2*10^{-16}$	Yes	1.548

Residual Deviance: 25467 on 26545 degrees of freedom

2. **Model selected by AIC in backwards direction**

$$\log\left(\frac{E(Y|\mathbf{X})}{1-E(Y|\mathbf{X})}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6$$

	Coefficient estimate	Std.Error	z value	p-value	Significant?	VIF
Intercept	-21.389 (β_0)	0.300	-71.279	$2*10^{-16}$	Yes	

Points of home team(X1)	-0.014 (β_1)	0.002	-7.167	$7.69*10^{-13}$	Yes	1.979
Field Goal percentage of home team(X2)	26.037 (β_2)	0.494	52.765	$2*10^{-16}$	Yes	2.211
Free throw percentage of home team(X3)	3.397 (β_3)	0.164	20.710	$2*10^{-16}$	Yes	1.126
4-points percentage of home team (X4)	3.940 (β_4)	0.165	23.941	$2*10^{-16}$	Yes	1.181
#Assists of home team(X5)	0.002 (β_5)	0.004	0.405	0.685	No	1.441
#Rebounds of home team (X6)	0.169 (β_6)	0.003	55.033	$2*10^{-16}$	Yes	1.548

Residual Deviance: 25467 on 26545 degrees of freedom

3. Model selected by BIC in backwards direction

$$\log(\frac{E(Y|\mathbf{X})}{1-E(Y|\mathbf{X})}) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_6X_6$$

	Coefficient estimate	Std.Error	z value	p-value	Significant?	VIF
Intercept	-21.395 (β_0)	0.300	-71.437	2*10-16	Yes	
Points of home team(X1)	-0.014 (β_1)	0.002	-7.411	1.25*10-13	Yes	1.786
Field Goal percentage of home team (X2)	26.076 (β_2)	0.484	53.838	2*10-16	Yes	2.129
Free throw percentage of home team(X3)	3.392 (β_3)	0.164	20.736	2*10-16	Yes	1.120
3-points percentage of home team(X4)	3.946 (β_4)	0.164	24.061	2*10-16	Yes	1.172
#Rebounds of home team(X6)	0.170 (β_6)	0.003	55.111	2*10-16	Yes	1.544

Residual Deviance: 25467 on 26546 degrees of freedom

4. Model selected by LASSO

$$\log(\frac{E(Y|\mathbf{X})}{1-E(Y|\mathbf{X})}) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6$$

	Coefficient estimate	Std.Error	z value	p-value	Significant?	VIF
Intercept	-21.389 (β_0)	0.300	-71.279	$2*10^{-16}$	Yes	
Points of home team(X1)	-0.014 (β_1)	0.002	-7.167	$7.69*10^{-13}$	Yes	1.979
Field Goal percentage of home team(X2)	26.037 (β_2)	0.494	52.765	$2*10^{-16}$	Yes	2.211
Free throw percentage of home team(X3)	3.397 (β_3)	0.164	20.710	$2*10^{-16}$	Yes	1.126
5-points percentage of home team (X4)	3.940 (β_4)	0.165	23.941	$2*10^{-16}$	Yes	1.181
#Assists of home team(X5)	0.002 (β_5)	0.004	0.405	0.685	No	1.441
#Rebounds of home team (X6)	0.169 (β_6)	0.003	55.033	$2*10^{-16}$	Yes	1.548

Residual Deviance: 25467 on 26546 degrees of freedom

5. Likelihood ratio test:

H0: $\beta_5 = 0$; Ha: $\beta_5 \neq 0$

Test Statistics: 0.164 ~ chi-square distribution with degree freedom=1

P-value: 0.685

Fail to reject H0, so we conclude $\beta_5 = 0$

6. Final model = BIC_Model

$$\log(\frac{E(Y|\mathbf{X})}{1-E(Y|\mathbf{X})}) = - 21.395 -0.014*\text{Points} + 26.076*\text{Field Goal Percentage} + 3.392*\text{Free Throw Percentage} + 3.946*\text{three points percentage} + 0.170*\text{Number of rebounds}$$

- Y ~ Bernoulli(p=E(Y|**X**)), where E(Y|**X**) is the probability that Y=1, or the home team wins, given **X**

Model Diagnostic

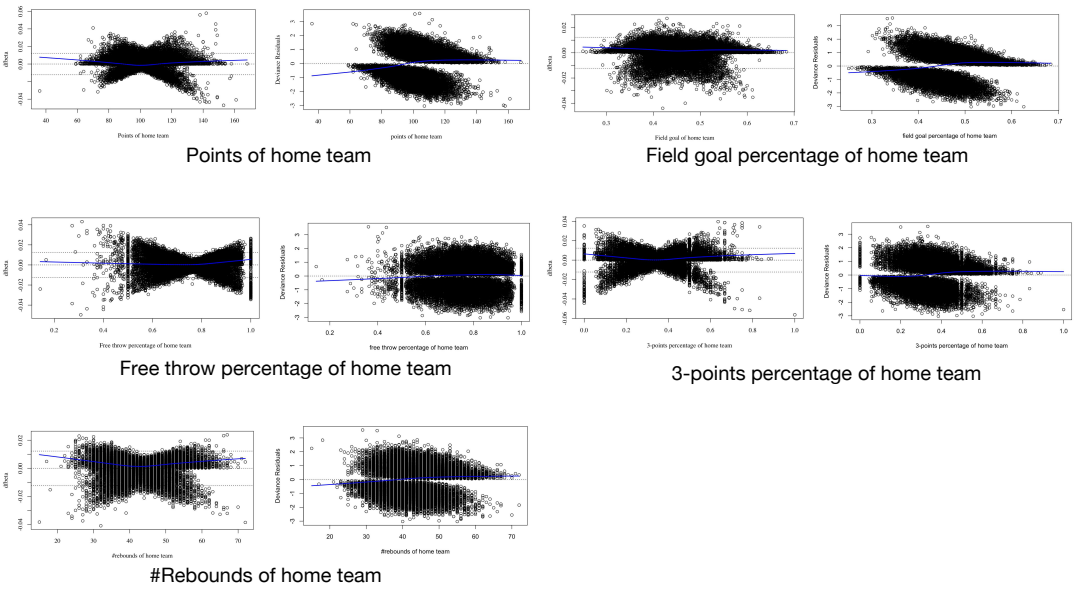


Figure 2: scatter plot of dfbeta and deviance residual of 5 predictors.

Firstly, the cutoff $\frac{-2}{\sqrt{n}}$ and $\frac{2}{\sqrt{n}}$ are used to identify the influential observations on the coefficient estimates. Every scatter plot of dfbeta and predictors show a X-shaped pattern. For example, the vertical spread around '3-points percentage' ≈ 0.35 is the smallest, and get larger in both direction and then eventually exceed the cutoff dashed line. This indicates the observations with '3-points percentage' ≈ 0.35 has less influence on the coefficient estimate compared to other numbers.

Secondly, in all deviance residuals plots, the blue LOWESS line is centred about the zero line, and there are slight cluster patterns and distant points but no extreme dispersions. This indicates the independence assumption may not hold perfectly, and the outliers exist. Interestingly, the location of smallest residual in all plots shows the model predicts more accurately at extreme value of predictors.

Thirdly, all VIF are less than 5, which indicates no presence of multicollinearity.

Further investigation on these problems is required, but we include them as limitations.

Model Validation

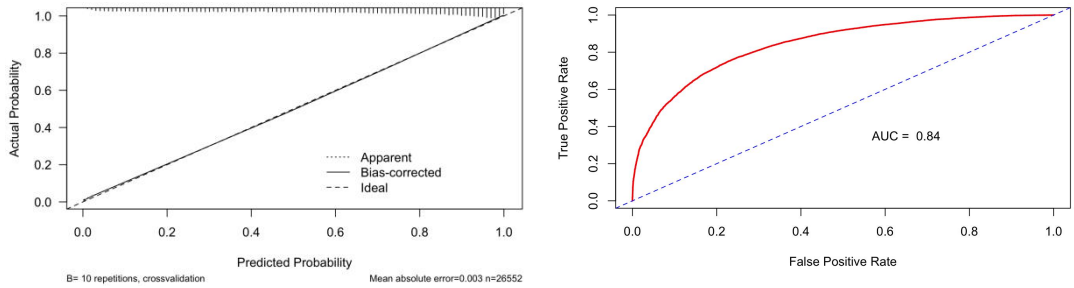


Figure 3:Calibration Plot and ROC curve

From the calibration plot, the bias corrected line nearly overlap the diagonal line which represents the predicted probability and actual probability are identical. Hence, final model demonstrates high accuracy without overfitting.

From the ROC curve plot, where x-axis it the true positive rate and y-axis is the true negative rate, the AUC is 0.84. This indicates a actual positive instance will be predicted as positive by the model 84% time.

4. Discussion

$$\log\left(\frac{E(Y|\mathbf{X})}{1-E(Y|\mathbf{X})}\right) = -21.395 - 0.014*\text{Points} + 26.076*\text{field goal Percentage} + 3.392*\text{free throw Percentage} + 3.946*\text{three points percentage} + 0.170*\text{Number of rebounds}$$

The coefficient of 'points' is negative while others are positive. For example, the odds of wining the game, is multiplied by $e^{-0.014}(\approx 0.98)$, when the 'points' increase by 1 and other predictors stay unchanged. Odds in this context is the ratio of winning probability to the lose probability. Therefore, the odds ratio less than 1 indicates victory becomes less likely as 'points' increases. Conversely, increase in other predictors will increase the winning probability. With this model, we are able to disclose the numerical relationship between the winning probability and performance statistics to make better game strategy, which answers the research question.

For limitations, the presence of influential observation may contains outlier or data entry errors that negatively impact the goodness of fit, instead of valid leveraged observations. This requires further investigation to decide whether to remove them. Also, the slightly violated independence assumption will lower the liability of inference. This can be corrected by transformation on the predictors.

Limitation
1.Problem with final model
2.Impact of problem
3.Why/how they could not be corrected

Reference List:

[1]Statista. Total league revenue of the NBA since 2005. Statista. https://www.statista.com/statistics/193467/total-league-revenue-of-the-nba-since-2005/

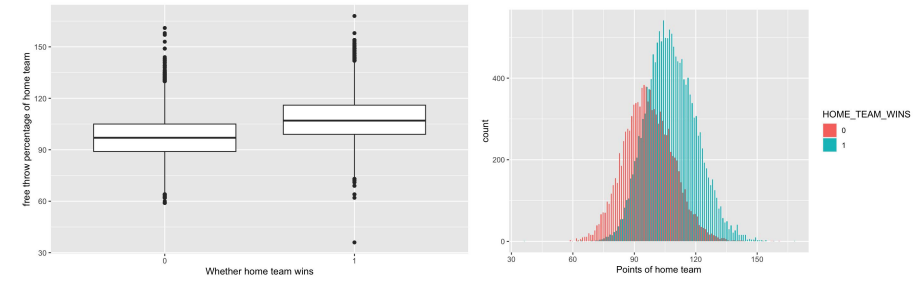
[2]Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA Game Result Prediction Using Feature Analysis and Machine Learning. Annals of Data Science, 6(1), 103–116. https://doi.org/10.1007/s40745-018-00189-x

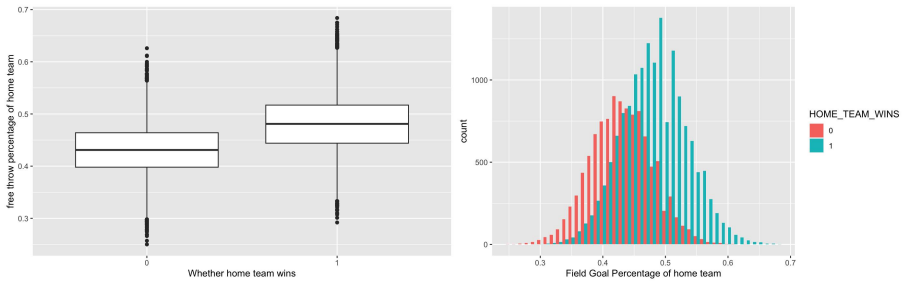
[3]Song, K., Zou, Q., & Shi, J. (2020). Modelling the scores and performance statistics of NBA basketball games. Communications in Statistics. Simulation and Computation, 49(10), 2604–2616. https://doi.org/10.1080/03610918.2018.1520878

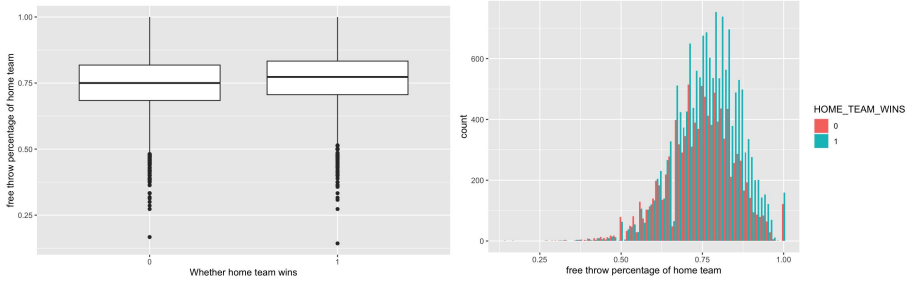
[4]McGoldrick, K., & Voeks, L. (2005). We Got Game!: An Analysis of Win/Loss Probability and Efficiency Differences Between the NBA and WNBA. Journal of Sports Economics, 6(1), 5–23. https://doi.org/10.1177/1527002503262649

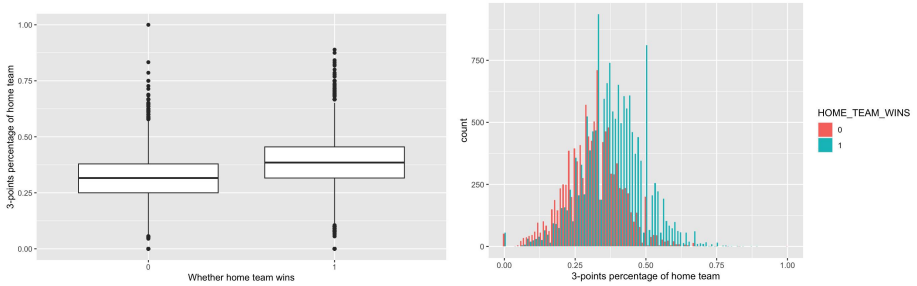
[5]Lauga, N. 2023. NBA Games, Kaggle. https://www.kaggle.com/datasets/nathanlauga/nba-games?select=games.cs

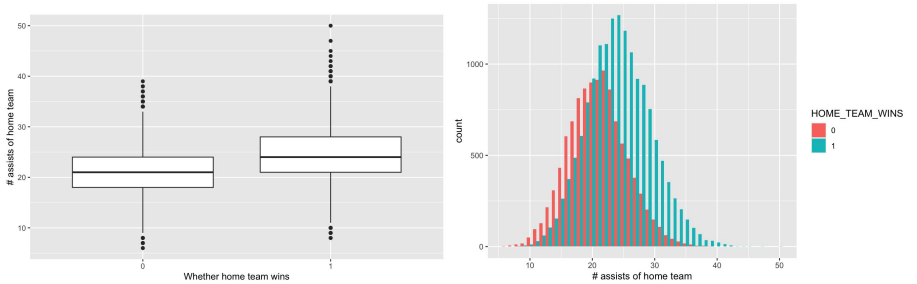
Appendix A: Histogram and boxplots of 6 predictors classified by the response=0 and 1. For example, the distribution of 'points of home team' for wins is higher than that for losses, which indicates this predictor is potentially significant.

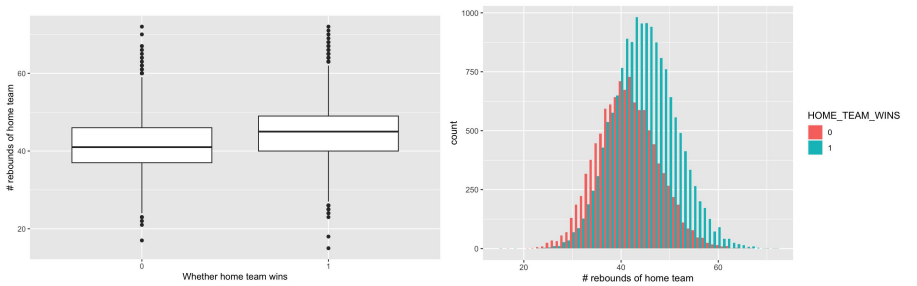












Appendix B. Scatter plots of predicted probability and predictors. The linear LOWESS trend in every plot indicates the linearity assumption is met.