NBA REGRESSION PROJECT

ADI IYER, ELIE TCHUMA, ASHWIN KURUCHI, ISHAN LAMBA

PROBLEM AND CONTEXT

- **Context and Importance**
- NBA teams and coaches would like to know:
  - Which in game factors influence victories
  - What **weight** each factor has on influencing victories
  - Which in game factors **interact** with each other
- **Questions**
  - What factors can be **most relevant in** predict if a home NBA team wins a game?
  - Can we build a model that accurately predicts if a team wins the game?

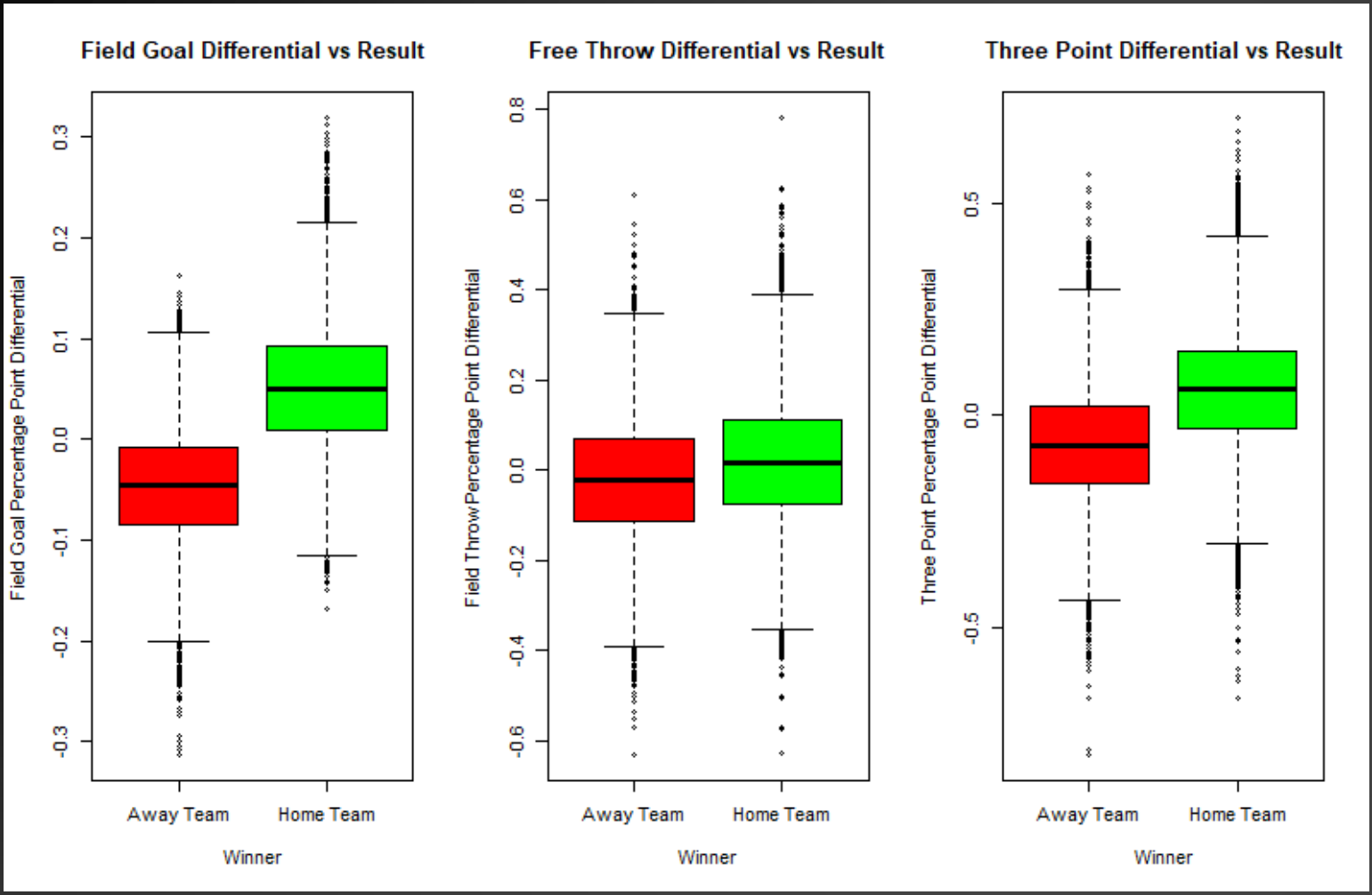
DATA FOR THIS PROJECT

- Large (>25000) number of 2021 box scores for NBA games
- **Binomial response (**1 if home team wins, 0 if home team loses)
- **Predictors of interest**
  - Field goal percentage differential (home-away)
  - Free throw percentage differential (home-away)
  - Three point percentage differential (home-away)
  - Assist differential (home-away)
  - Rebound differential (home-away)

MODEL BUILDING DECISIONS

- Chose to utilize predictors differentials (home-away) to simplify analysis
- Chose to utilize **logistic regression** since response was binomial
- Chose to utilize **out of sample prediction (90% training data, 10% test data)** to evaluate model efficacy
- **Steps in our analysis**
- Exploratory data analysis → Logistic regression model building → Out of sample prediction for testing
- **Conclusion =** Each predictor was important for the model.

EXPLORATORY DATA ANALYSIS 1



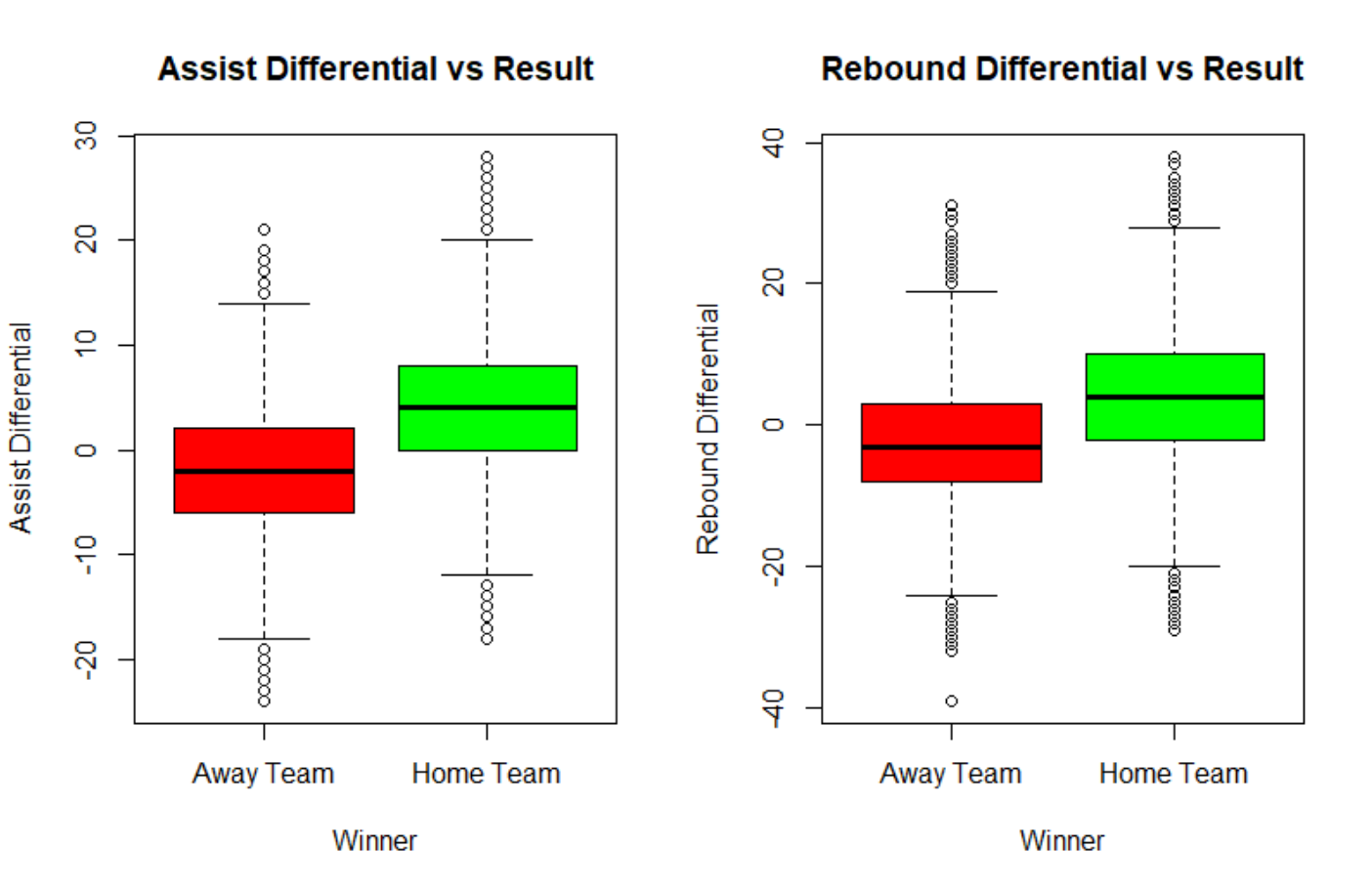
Boxplots for field goal, free throw, and three point percentages against victories

Not many obvious outliers requiring extra investigation

Predictor spread across each group (win or loss) appears to be similar for each predictor

In each case, the predictor **mean** is higher for win than for loss. Significance is debatable

EXPLORATORY DATA ANALYSIS 2

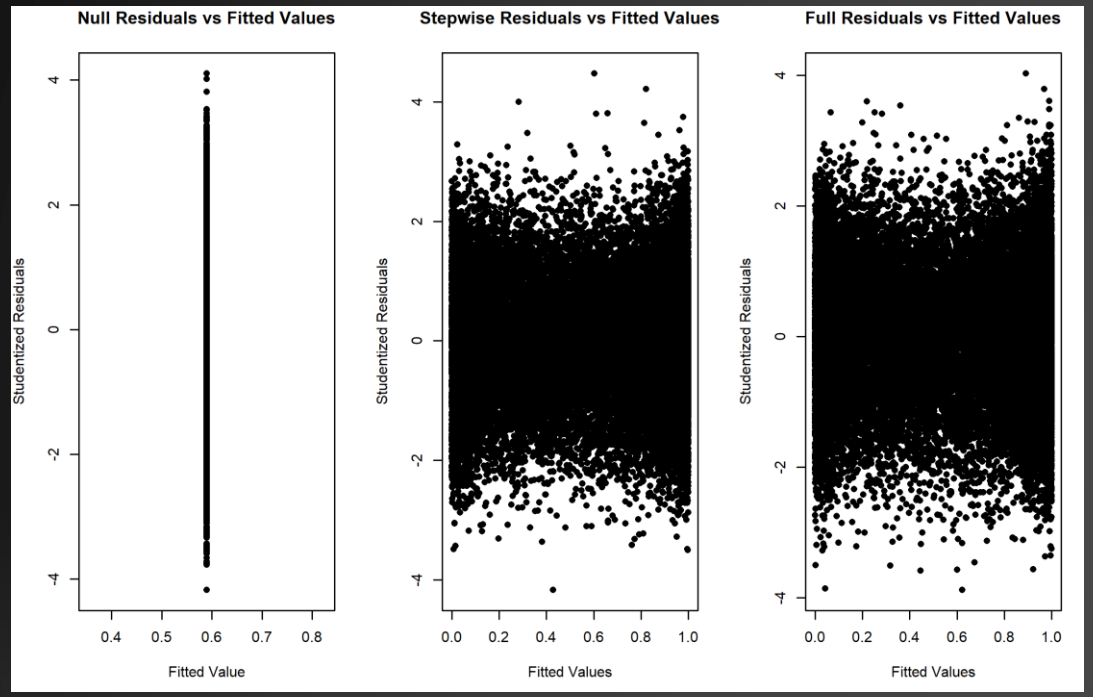


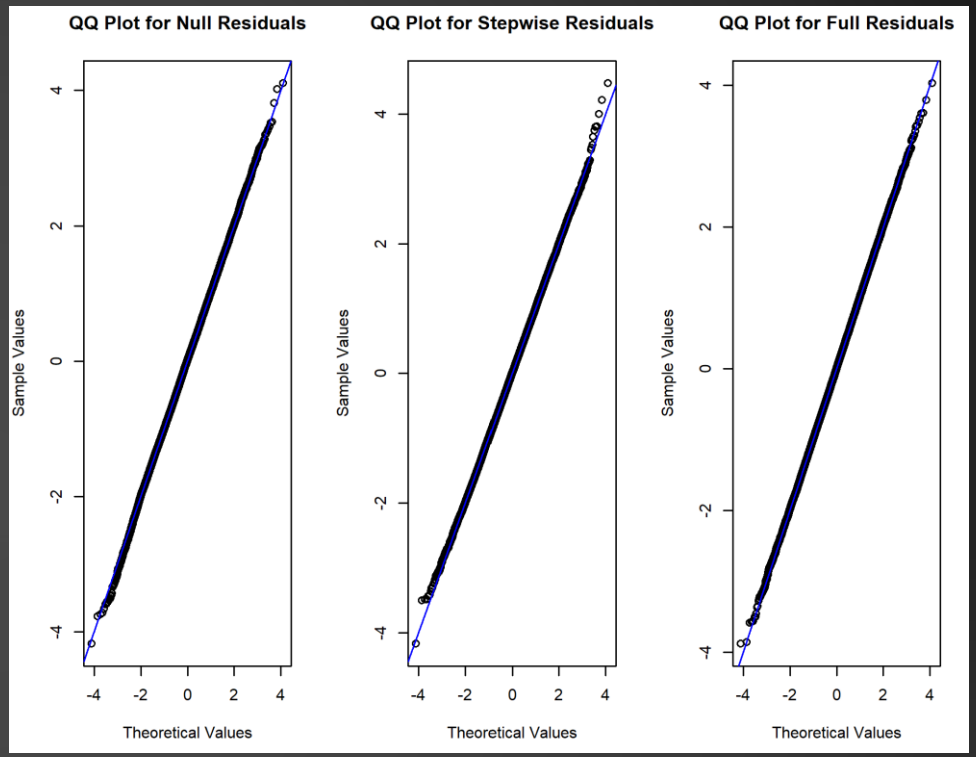
Boxplots for assist and rebound differentials against victories

Again, outliers for each predictor seem to be limited

More **average** assists and rebounds for home team when they win against when they lose

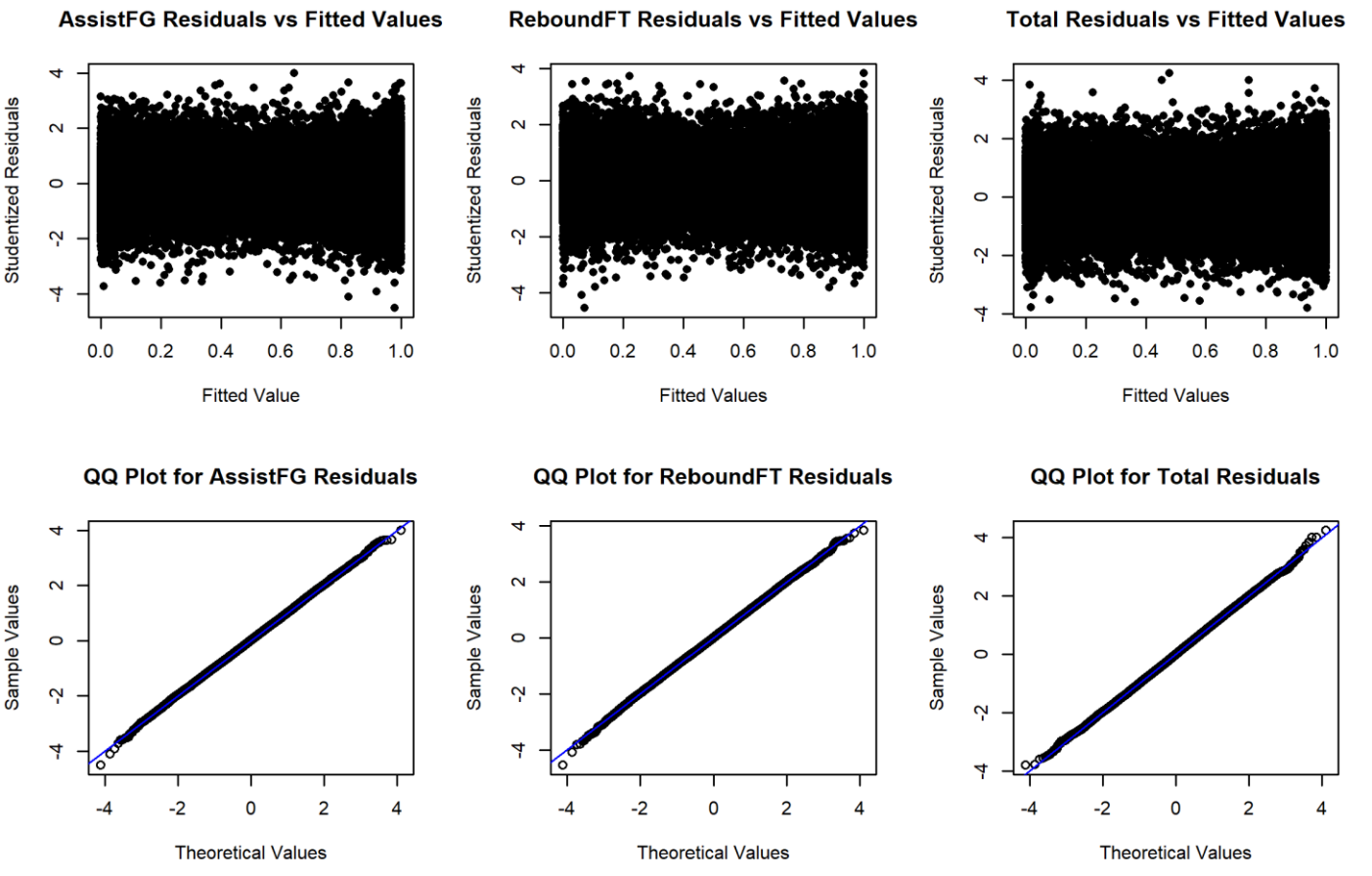
STEPWISE MODEL BUILDING DIAGNOSTICS





- Evaluated a null model, stepwise model (AIC), and full model
- Assumptions for the residuals for these models appear to be OK

INTERACTION MODEL BUILDING DIAGNOSTICS



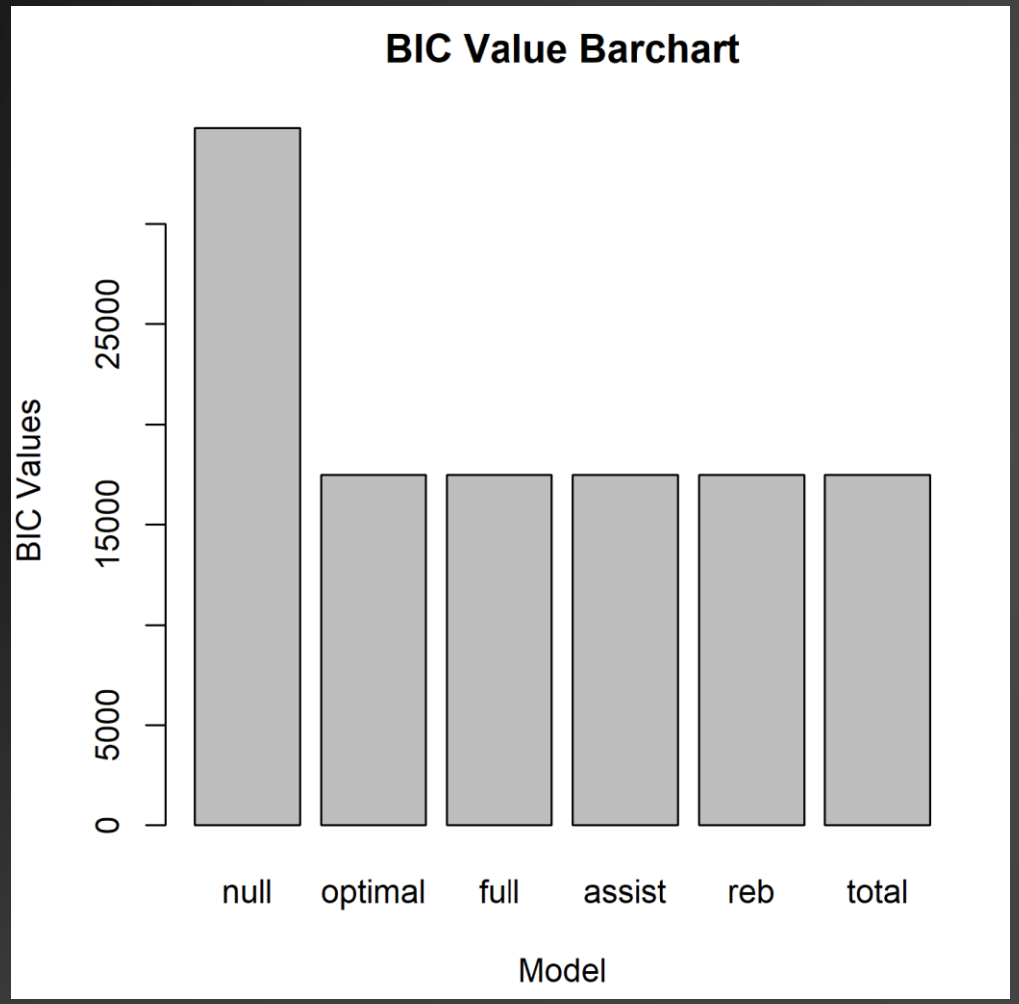
Evaluated model with assist and field goals interaction (more assists lead to high quality shots)

Evaluated model with rebound and free throw interaction (more rebounds, more likelihood to be fouled)

Evaluated total model including these interactions

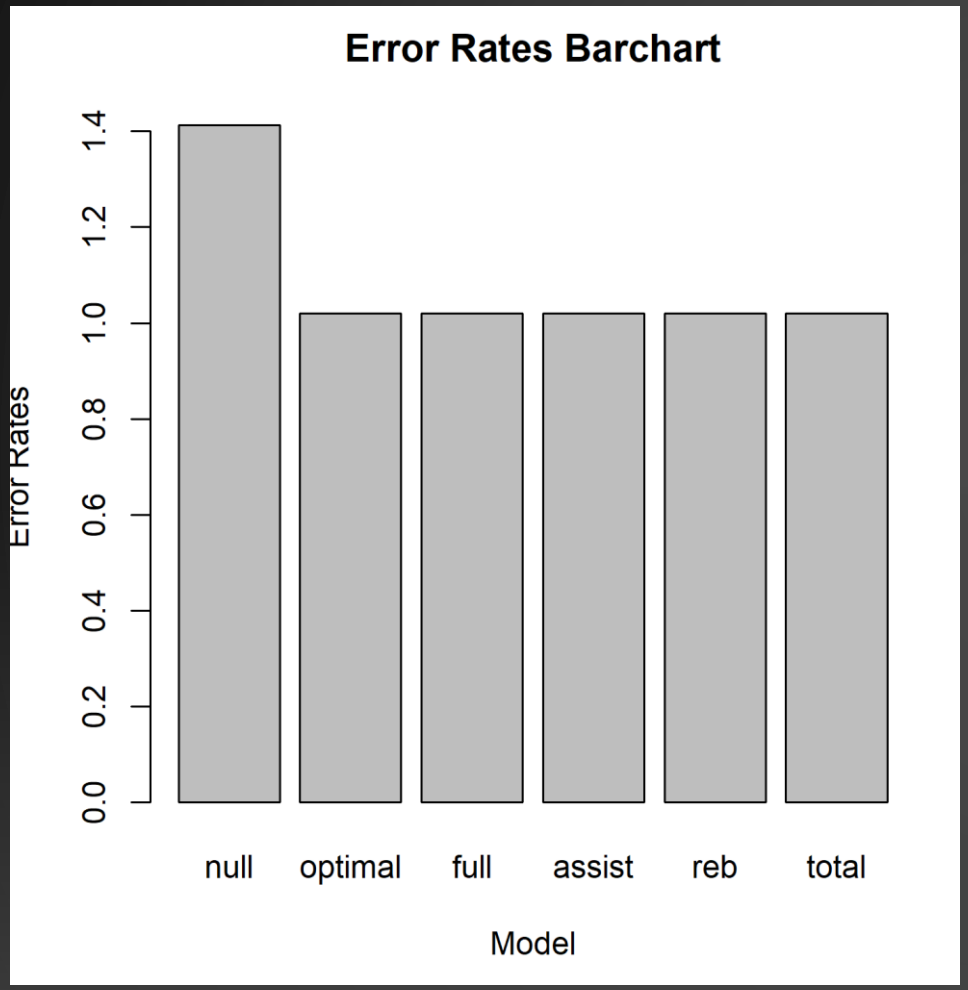
Assumptions look OK. Can proceed with this model

EVALUATION TECHNIQUE 1



- We utilized BIC to evaluate the model.
- The **optimal model** utilizing forward selection demonstrated that each of the differentials was highly significant (p value < 0.0001).
- Interactions had practically no effect on improving the model itself
- Model with each of the predictors appeared to be most useful

EVALUATION TECHNIQUE 2



- We utilized out of sample prediction to evaluate the model
- The model did not seem to be good at predicting whether or not a team wins games or not
- The model seemed to get many incorrect values when checking.
- We think this could be due to insufficient predictors

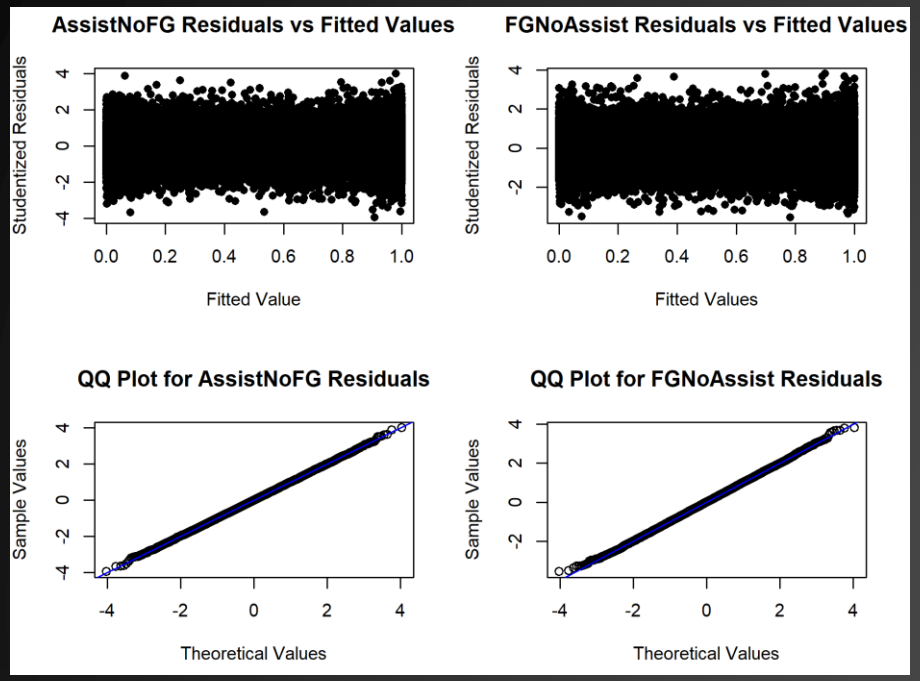
ASSESSMENT OF MULTICOLLINEARITY

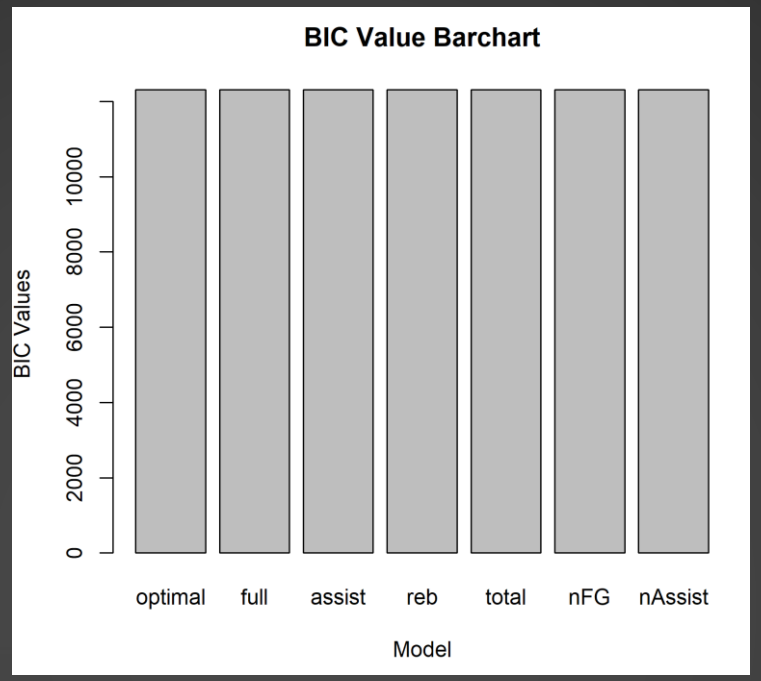
Multicollinearity was assessed via correlation plots and variance inflation factors (VIFs)

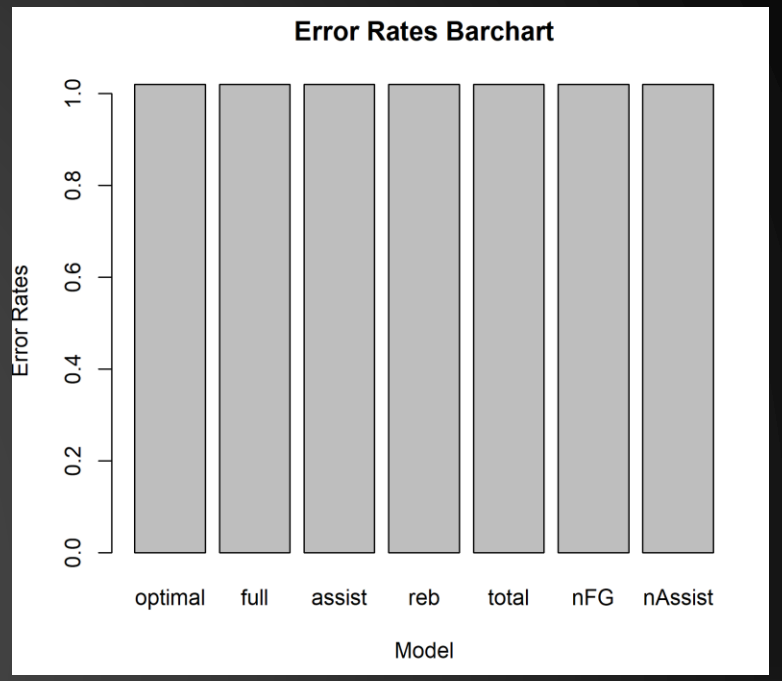
Variance inflation factors (VIFs) were a bit larger than one

Analyzing the correlation plot, it seems that the assist differential and the field goal percentage differential appear to be somewhat correlated

Models were built that had only assists differential and only field goal percentage differential







While the interaction models seemed good (diagnostics), they did not assist in reducing BIC and increasing prediction accuracy

Multicollinearity does not appear to have a major role in affecting model accuracy

HIGHLIGHTS

- Each of the models were effective at achieving diagnostics
- Each of the predictors was significant, thus the **optimal and the full model** had the same predictors
- Interactions and multicollinearity had practically zero affect
- Model accuracy was limited, and more data on distinct predictors would be good
- **Recommendation = We would utilize the optimal model because it has the lowest amount of predictors and virtually identical BIC and prediction accuracy**