Final Project Part 3

Introduction

Measuring a player's all-around impact on their basketball team is a difficult but important task. Fundamentally basketball games are won by one team outscoring the other team throughout the duration of the game. Research has show that teams with high points differential Therefore, a reasonable measure of player's impact is box-plus-minus (BPM) which is a per-game statistic that measures the points differential of the player's team and the opposing team when the player is playing in the game. Previous research has concluded that BPM is a useful metric for evaluating a player's value (Terner 2021). The goal of this research project is to study how a player's individual game performance contributes to overall team success. In this research project, a linear regression model will be created to predict a player's average BPM based on their per-game stats and understand which factors influence the statistic. By being able to understand what factors influence a player's BPM team managers are able to construct a team and allocate their resources in a way which optimizes team success. Furthermore, by being able to accurately predict a player's BPM the team managers will have a rough idea of how impactful future prospects will be for their team.

Methods

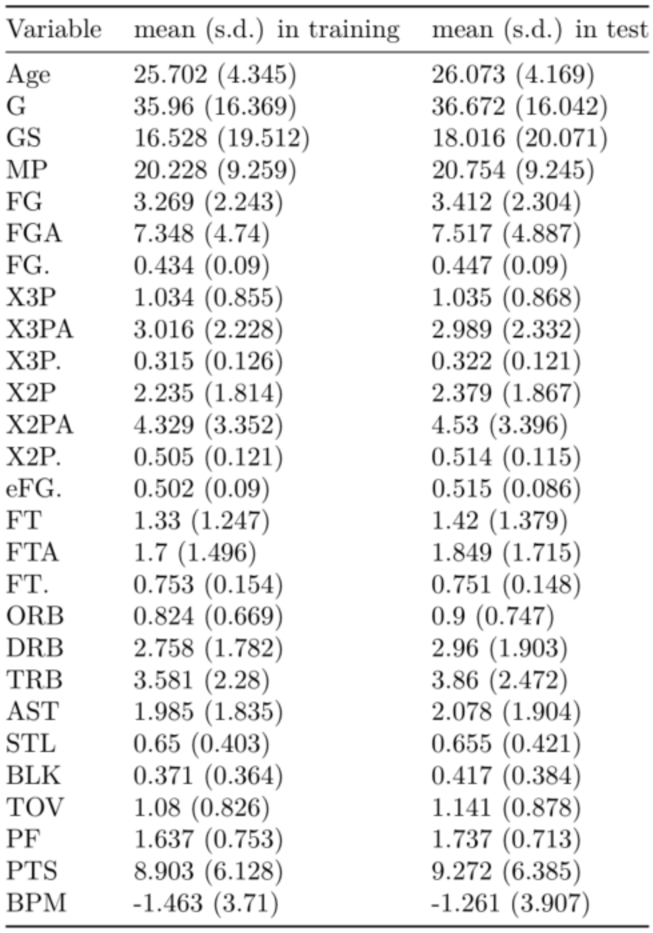
Initially in the analysis, the dataset is equally split into two parts, a training dataset and test dataset. The training dataset will be used to build and decide on a final model. Statistical tools used to find a final model are anova F-test, variance-inflation factor (vif), adjusted R-squared values, all possible subsets selection method and stepwise selection method. The anova F-test is used to verify whether at-least one predictor in a given model is significant to the response (BPM). If a model passes the anova F-test it should provide a good starting point for constructing a final model. The vif is used to detect possible multicollinearity in the predictors which should be done before constructing a model but after checking assumptions of a full model. Vif is used to decide which predictors in model have multicollinearity. Based on severity of the vif a decision will be made for whether or not those predictors should be removed from the model. The all-subsets selection method is used to find the best model for each number of predictors based on adjusted R-squared values. Then the best models of each size can be compared using their adjusted R-squared values to decide on model that has a balance of explaining a substantial amount of variation in the response as well as simplicity for comprehension. The stepwise selection method is one in which a model is chosen by adding or removing one predictor at a time based on a penalty term such as AIC or BIC until predictors cannot be added nor removed. This will result in a model that accounts for the conditional nature of regression and can be compared to the all-possible subsets method to help decide on a final model. After each model is created, conditions 1 and 2 are checked to ensure that residual plots of the model can be interpreted correctly. The residual plots will then be checked for any violations of the linearity, non-constant variance, uncorrelated errors, and normality assumptions.

If the linearity, non-constant variance, or normality assumptions are violated a transformation can be used to attempt to correct the assumption violation. However, if the uncorrelated errors assumption is violated it should be discussed as a limitation of the model. Additionally, problematic observations such leverage, outliers, and influential points should be checked for each model and should be used to compare different models. If problematic observations are found they should be discussed as limitations of the model and can be used as justification as to why one model is superior to another. Once decided on a final or between a few models, the test dataset will be used to validate the model(s) to conclude whether a model is a good model or has been overfitted on the training dataset. Assumptions and problematic observations should once again be checked when doing model diagnostics on models using the test dataset.

Results

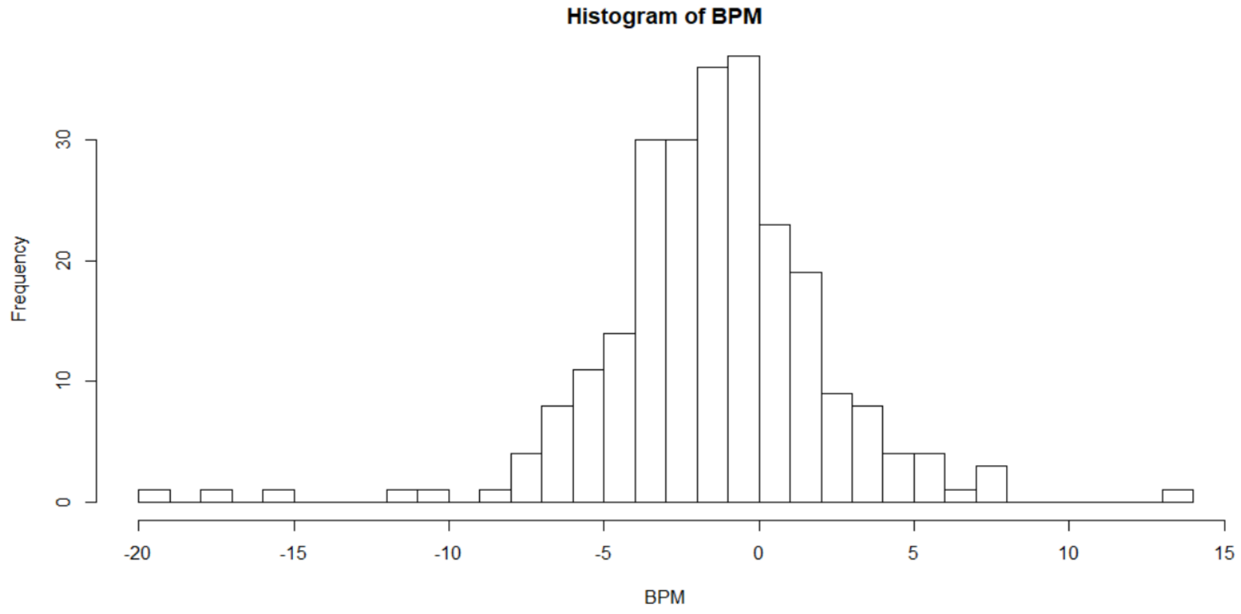
Data Summary:

Table 1: Summary statistics in training and test dataset



Variable	mean (s.d.) in training	mean (s.d.) in test
Age	25.702 (4.345)	26.073 (4.169)
G	35.96 (16.369)	36.672 (16.042)
GS	16.528 (19.512)	18.016 (20.071)
MP	20.228 (9.259)	20.754 (9.245)
FG	3.269 (2.243)	3.412 (2.304)
FGA	7.348 (4.74)	7.517 (4.887)
FG.	0.434 (0.09)	0.447 (0.09)
X3P	1.034 (0.855)	1.035 (0.868)
X3PA	3.016 (2.228)	2.989 (2.332)
X3P.	0.315 (0.126)	0.322 (0.121)
X2P	2.235 (1.814)	2.379 (1.867)
X2PA	4.329 (3.352)	4.53 (3.396)
X2P.	0.505 (0.121)	0.514 (0.115)
eFG.	0.502 (0.09)	0.515 (0.086)
FT	1.33 (1.247)	1.42 (1.379)
FTA	1.7 (1.496)	1.849 (1.715)
FT.	0.753 (0.154)	0.751 (0.148)
ORB	0.824 (0.669)	0.9 (0.747)
DRB	2.758 (1.782)	2.96 (1.903)
TRB	3.581 (2.28)	3.86 (2.472)
AST	1.985 (1.835)	2.078 (1.904)
STL	0.65 (0.403)	0.655 (0.421)
BLK	0.371 (0.364)	0.417 (0.384)
TOV	1.08 (0.826)	1.141 (0.878)
PF	1.637 (0.753)	1.737 (0.713)
PTS	8.903 (6.128)	9.272 (6.385)
BPM	-1.463 (3.71)	-1.261 (3.907)

From looking at Table 1 we see that the training and test dataset are very similar in relation to the summary of each variable.



We can see from the histogram of the response (BPM) that it is normally distributed.

Analysis Process and Results

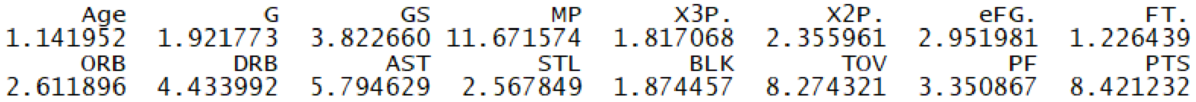
I began constructing a final model by creating a full model which includes every predictor. I used the anova F-test to determine whether the full model has at least one significant predictor and it passed. I then checked the conditions as well as assumptions for the model and found that that all were held. I proceeded by checking for multicollinearity in the model using the vif of each predictor. I found that some predictors had extremely high vif values causing a multicollinearity problem.

Variance Inflation Factor (vif) for each numeric predictor in the full model:

```
          Age             G            GS            MP            FG           FGA
     1.250763      2.068105      4.262793     13.927427   5424.750742  11652.514605
          FG.           X3P          X3PA          X3P.           X2P          X2PA
    26.177365    624.735363   2509.542464      2.528928   1663.141022   5865.256058
         X2P.          eFG.            FT           FTA           FT.           ORB
     3.073304     25.120639    403.904587    102.424026      1.896208    211.548189
          DRB           TRB           AST           STL           BLK           TOV
  1550.716034   2512.281871      6.339453      2.613384      1.932085      9.031014
           PF           PTS
     3.498963   8401.159635
```

To reduce the multicollinearity present in the model, I attempted to remove predictors that were mostly summarized by other predictors. For example, “X3P.” (three-point percentage) summarizes “X3P” (three pointers made) and “X3PA” (three pointer attempts), while I feel that taking “X3PA” into account along with the “X3P.” is crucial, by removing those predictors we get significantly less multicollinearity in our model. Similar reasoning was used to justify removing “FG”, “FGA”, “FG.”, “FT”, and “FTA”. Furthermore, “TRB” was removed as “TRB” = “ORB” + “DRB”, therefore “TRB” can be removed to reduce multicollinearity while not losing any additional information. Additional, “eFG.” (effective field goal percentage) is the number of shots made divided by the number of shot attempts, however “X2P” has a weight of 1 and “X3P” have weight of 1.5. Therefore, “eFG.” summarizes FG. (field goal percentage) and gives back some information that was lost when removing X3P and X3PA. While there are some predictors that still have concerning VIF values to worry about such as MP, AST, TOV, and PTS, all these predictors should be highly related to the response and reflect a player's impact on team success. After removing some predictors, multicollinearity in the model significantly reduced.

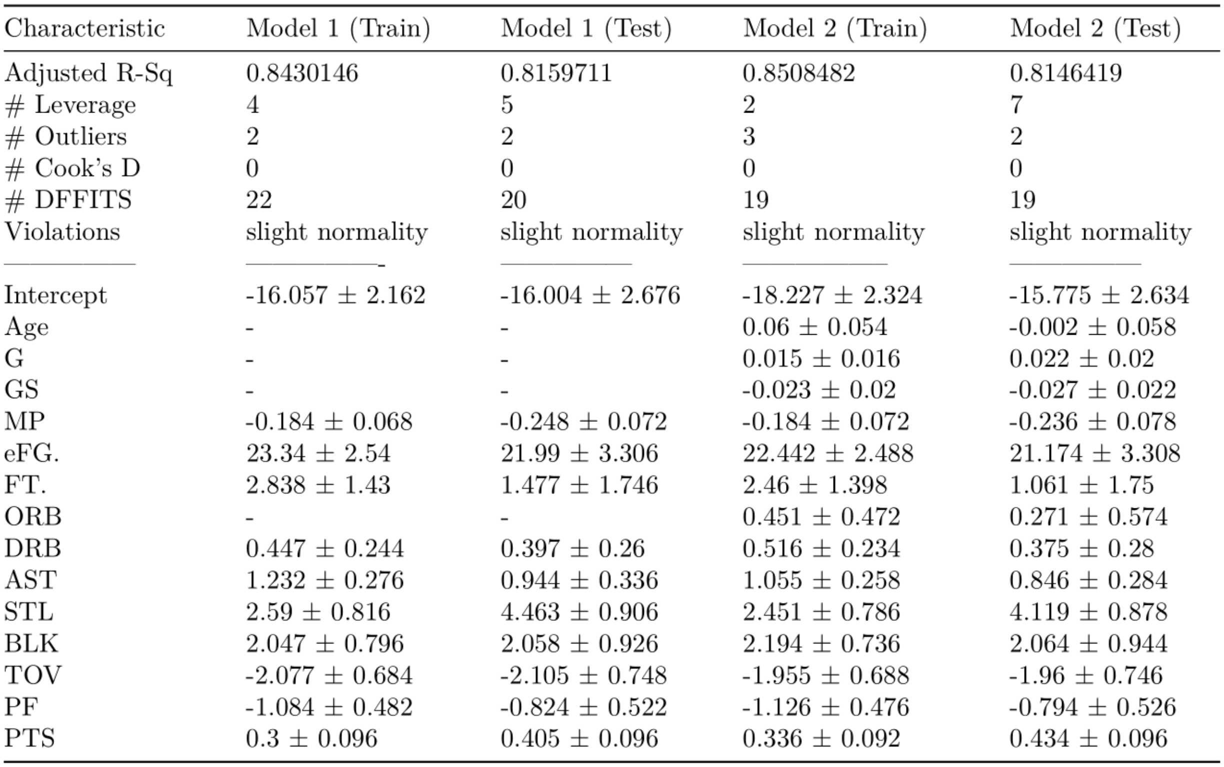
Variance Inflation Factor (vif) for each predictor in the full model after reducing multicollinearity:



```
      Age         G        GS        MP      X3P.      X2P.      eFG.       FT. 
 1.141952  1.921773  3.822660 11.671574  1.817068  2.355961  2.951981  1.226439 
      ORB       DRB       AST       STL       BLK       TOV        PF       PTS 
 2.611896  4.433992  5.794629  2.567849  1.874457  8.274321  3.350867  8.421232 
```

Therefore, moving forward I considered only the predictors above for the final model to significantly reduce multicollinearity. Then I used the all-possible subsets method for insight on what predictors should be included in the final model. I found that models with 10 or greater predictors had very negligible adjusted R-squared values and decided that the model with 10 predictors had the best balance of simplicity and high adjusted R-squared value. For comparison, I decided to create another model using the stepwise selection method. I found that both methods were largely agreeing with the same predictors. Both models had no assumptions violations other than a minor normality violation.

Summary of Models (Model 1 = All Possible Subsets Model, Model 2 = Stepwise Selection Model)



Characteristic	Model 1 (Train)	Model 1 (Test)	Model 2 (Train)	Model 2 (Test)
Adjusted R-Sq	0.8430146	0.8159711	0.8508482	0.8146419
# Leverage	4	5	2	7
# Outliers	2	2	3	2
# Cook's D	0	0	0	0
# DFFITS	22	20	19	19
Violations	slight normality	slight normality	slight normality	slight normality
————	————-	————	————	————
Intercept	-16.057 ± 2.162	-16.004 ± 2.676	-18.227 ± 2.324	-15.775 ± 2.634
Age	-	-	0.06 ± 0.054	-0.002 ± 0.058
G	-	-	0.015 ± 0.016	0.022 ± 0.02
GS	-	-	-0.023 ± 0.02	-0.027 ± 0.022
MP	-0.184 ± 0.068	-0.248 ± 0.072	-0.184 ± 0.072	-0.236 ± 0.078
eFG.	23.34 ± 2.54	21.99 ± 3.306	22.442 ± 2.488	21.174 ± 3.308
FT.	2.838 ± 1.43	1.477 ± 1.746	2.46 ± 1.398	1.061 ± 1.75
ORB	-	-	0.451 ± 0.472	0.271 ± 0.574
DRB	0.447 ± 0.244	0.397 ± 0.26	0.516 ± 0.234	0.375 ± 0.28
AST	1.232 ± 0.276	0.944 ± 0.336	1.055 ± 0.258	0.846 ± 0.284
STL	2.59 ± 0.816	4.463 ± 0.906	2.451 ± 0.786	4.119 ± 0.878
BLK	2.047 ± 0.796	2.058 ± 0.926	2.194 ± 0.736	2.064 ± 0.944
TOV	-2.077 ± 0.684	-2.105 ± 0.748	-1.955 ± 0.688	-1.96 ± 0.746
PF	-1.084 ± 0.482	-0.824 ± 0.522	-1.126 ± 0.476	-0.794 ± 0.526
PTS	0.3 ± 0.096	0.405 ± 0.096	0.336 ± 0.092	0.434 ± 0.096

Note: Indicator variables were not included in this summary as there were too many

Both models performed very similarly with the test dataset. However, Model 1 provides a very comparable adjusted R-value while also having less a few less predictors. For this reason, I decided to choose Model 1 as my final model.

Discussion

From the final model we see that our analysis has concluded that MP, eFG., FT., DRB, AST, STL, BLK, TOV, PF, PTS, and Tm are a player's per-game stats that most influence their BPM. It makes sense to include these predictors as they reflect what actions the player did to contribute to their team. I find it fascinating how much larger the coefficient of eFG. than every other predictor. The coefficient of eFG. in our final model using the training dataset tells us on average a player's box-plus-minus increases by 23.34 when that player's effective field goal percentage increases by 1% and all other predictors are held constant. While this does seem quite high, it seems reasonable that as a player shoots better on average their team's points differential will be higher. This model can be used to evaluate player contribution and understand which skills in basketball are most important for team success. Furthermore, scouts can use this model to predict the value a future prospect brings to their team.

Limitations

A limitation of the model is that basketball is a team sport and therefore much of a player's BPM is influenced by their teammates. For example, a poorly skilled player that plays on a team with very good players will likely have a higher BPM simply because they're team is very good and outscores their opponents while that player is playing. To account for this a team predictor was included in the final model. Another limitation is that many per-game stats are related to each other. Therefore, some multicollinearity will always exist in the model due to the nature of the predictors. Additionally, there were a few problematic observations which may influence the accuracy of the model.

References

Grassetti, Bellio, R., Di Gaspero, L., Fonseca, G., & Vidoni, P. (2021). An extended regularized adjusted plus-minus analysis for lineup management in basketball using play-by-play data. IMA Journal of Management Mathematics, 32(4), 385-409. https://doi.org/10.1093/imaman/dpaa022

Deshpande, & Jensen, S. T. (2016). Estimating an NBA player's impact on his team's chances of winning. Journal of Quantitative Analysis in Sports, 12(2), 51-72. https://doi.org/10.1515/jqas-2015-0027

Terner, & Franks, A. (2021). Modeling Player and Team Performance in Basketball. Annual Review of Statistics and Its Application, 8(1), 1-23. https://doi.org/10.1146/annurev-statistics-040720-015536