My title*

My subtitle if needed

Yihang Cai

April 7, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

You can and should cross-reference sections and sub-sections. We use R Core Team (2023) and (**rohan?**).

The remainder of this paper is structured as follows. Section 2....

2 Data

The dataset is obtained from Basketball Reference Sports Reference LLC (2024), which contains NBA players' statistics in the 2023 to 2024 season. The dataset is directly downloaded from Basketball Reference using the given instruction. The dataset contains different statistics of NBA players, for instance players' position, age, assists, steals, etc. There are in total of 718 observations before any data cleanings.

Table 1: Top 5 NBA Players Based on Select Predictors Highly Correlated with Points Scored, Season 2023-2024

Player	3-point goals	2-point goals	free throws	points
Precious Achiuwa	25	204	69	552
Precious Achiuwa	13	65	24	193
Precious Achiuwa	12	139	45	359
Bam Adebayo	10	470	276	1246

*Code and data are available at: https://github.com/peachvegetable/NBA-player-points

Player	3-point goals	2-point goals	free throws	points
Ochai Agbaji	61	103	28	417

This dataset contains player statistics including 3-point goals, 2-point goals, and free throws, points can be calculated solely using these variables. My goal is to predict the number of scores a NBA player scores based on his performance, which means that these features are too closely related to the target variable (i.e. points). Table 1 shows the table with closely related variables, we can calculate the points a player scored directly using these variables. For instance, player Bam Adebayo had 10 3-points goals, 470 2-point goals and 276 free throws during the 2023-2024 season, so his total points gained is $10 \cdot 3 + 470 \cdot 2 + 1 \cdot 276 = 1246$, which is exactly the points he scored. Thus, those variables are removed from the dataset using 'tidyverse' package in R Wickham et al. (2019). Moreover, there are 12 different positions in the dataset, I categorized them into a broader classification which has only three categories - Guards (G): SG, PG, SG-PG, PG-SG, Forwards (F): SF, PF, PF-SF, SF-PF, SF-SG, Centers (C): C, PF-C, C-PF. This is done so that I don't need to create overwhelming dummy variables when generating the model and this process of binning could simplify the model and potentially strengthen the signal by reducing noise and multicollinearity.

3 Model

Lasso regression is a variation of linear regressions that can actually perform feature selection by setting the coefficients of less important features to zero. It has an additional parameter λ, which is the regularization parameter that controls the strength of the penalty. The penalty term shrinks some of the coefficients (i.e. the less important variables) toward zero. As λ increases, more coefficients are set to zero, leading to a simpler model. The value of λ is determined through cross-validation.

3.1 Model set-up

$$y_i = \beta_0 + \beta_i \cdot X_i \tag{1}$$

In this equation, y_i is the number of points a player scores, which is the dependent variable I am trying to predict. β_0 is the interception, and β_i is a matrix that contains the coefficients $\beta_1, \beta_2, ..., \beta_{18}$ for each predictor that the lasso regression will estimate. X_i is also a matrix contains the predictors: players position, age, games, game starts, minutes played, field goal percentage, 3-point field goal percentage, 2-point field goal percentage, effective field goal percentage, free throw percentage, offensive rebounds, defensive rebounds, total rebounds, assists, steals, blocks, turnovers, and personal fouls.

We run the model in R (R Core Team 2023) using the `tidymodels` package of Kuhn and Wickham (2020).

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

```
# #| include: false
# #| warning: false
# #| message: false
#
# first_lasso_model <- readRDS("../models/first_lasso_model.rds")
# second_lasso_model <- readRDS("../models/second_lasso_model.rds")
# predictions <- predict(second_lasso_model, test_data)
# results <- bind_cols(test_data, predictions)
# # Calculate RMSE
# rmse_results_2 <- rmse(results, truth = pts, estimate = .pred)
# rsq_results_2 <- rsq(results, truth = pts, estimate = .pred)
```

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

References

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.* https://www.tidymodels.org.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sports Reference LLC. 2024. "2023-2024 NBA Player Stats: Totals." Basketball-Reference.com. https://www.basketball-reference.com/leagues/NBA_2024_totals.html.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.