Forecasting NBA Players' Efficiency: Integrating College and Combine Predictors with Beta Regression

Areen Jain
Gretchen Whitney High School, Cerritos, CA

Abstract

This study uses beta regression and random intercept beta regression to model normalized Player Efficiency Rating (PER) in the NBA based on pre-draft data from college basketball and the NBA Combine. Using 463 players with complete statistics, we identify key predictors of efficiency, including position, weight, rebounds, field goal percentage, height, and vertical leap. The models account for the bounded nature of PER and capture longitudinal trends with player-specific random effects. Results highlight positional role as the strongest determinant, with guards and wings outperforming centers.

Keywords: beta regression, random intercept model, longitudinal analysis, player efficiency rating (PER), NBA draft prediction, college basketball, NBA Combine, positional roles, sports performance modeling

1 Introduction

1.1 Background

The National Basketball Association (NBA) is one of the most closely analyzed professional sports leagues, with teams, analysts, and researchers relying on box score metrics and advanced statistics to evaluate performance, project future success, and inform decisions on drafts, contracts, and roster construction. Predicting professional outcomes is challenging, as players must be evaluated in distinct but connected contexts – college basketball, the NBA Combine, and early NBA competition.

This study forecasts the Player Efficiency Rating (PER) in the NBA using college and NBA Combine data. Unlike traditional measures that emphasize scoring or isolated skills, PER incorporates all-around contributions derived from box-score statistics. By linking college performance, NBA pre-draft skills measurements (NBA Combine), and NBA early career outcomes, we identify which factors most effectively predict professional efficiency. This framework provides evidence-based information on player development and supports draft evaluation for front offices, scouts, and researchers.

On the offensive side, players generate value through their shooting efficiency, playmaking, ability to draw fouls, offensive rebounding, and avoidance of turnovers. Defensively, value is created through proper positioning, contesting and blocking shots, generating steals, rebounding, and limiting fouls. Because basketball is fluid and interconnected, a player's success cannot be defined by isolated actions. Instead, effectiveness comes from the ability to contribute across multiple phases of the game.

1.2 Literature Review

Predicting NBA draft success has moved from subjective judgments to data-driven models. Early approaches focused on physical traits measured at the NBA Combine. Teramoto et al. (2018) found that NBA Combine size measurements such as height, wingspan, and standing reach were strongly associated with defensive performance (Defensive Box Plus-Minus, DBPM; $r = 0.545$), highlighting the partial predictive value of physical data [1]. However, as Kannan et al. (2018) emphasized, advanced efficiency metrics such as field goal percentage (FG) and assists per game consistently outperformed raw physical statistics and totals in predicting NBA success, especially for guards and forwards [2].

Edwards et al. (2015) contributed with support vector machine (SVM) and principal component analysis (PCA) models and discovered that it is easier to predict whether a player will reach a basic threshold of success than to estimate their exact NBA Win Shares (WS). They concluded, in agreement with others, that historical performance and developmental context are more reliable than raw physical tests [3]. Similarly, Moxley and Towne (2015) challenged the value of "hidden potential" and found, using growth mixture models (GMMs), that college achievement and environment (rather than NBA Combine performance) predict which group a player will end up in, whether a role player or a potential star [4].

Greene (2015) bolstered this by showing that per-possession and rate-based metrics were superior for forecasting success, a theme echoed across recent studies [5].

Recent advancements have combined statistical analysis and subjective expert input. Mamonov (2023) used Extreme Gradient Boosting (XGBoost) to integrate college performance with mock draft rankings, demonstrating that this hybrid approach predicted player role tiers (Star, Above Average, Bench) with 75% accuracy, far better than using draft order alone [6]. Likewise, Kannan et al. (2018) highlighted that the combination of draft slot and college production improved F1-score from 0.54 (physical data alone) to 0.72 [2].

A recent leap is the Relevance-Based Prediction (RBP) method introduced by Czasonis et al. (2023), which selects historically similar players and dynamically adapts its predictions and features for each prospect, offering both individualized forecasts and clear reliability signals [7]. This method, like the broader trend, values transparency in explaining which data points and precedents shape the prediction.

1.3 Data Description

The dataset for this analysis was compiled from Kaggle and Basketball-Reference.com, combining collegiate, NBA Combine, and NBA performance metrics to predict long-term player success using a custom Player Efficiency Rating (PER), which summarizes overall on-court productivity. The initial pool included 3,840 NCAA Division I players from Kaggle, which was merged with NBA Combine-style physical and athletic data (height, weight, wingspan, vertical leap, sprint times) from Basketball-Reference.com, reducing the sample to 1,350 players with complete college and NBA Combine records. Filtering for players who appeared in at least one NBA regular-season game resulted in 692 players with full college, NBA Combine, and professional data. Significant missing values, particularly in NBA Combine metrics and advanced college stats, were addressed through manual cleaning and consolidation. Variables with over 30% missingness were excluded, reducing 53 original variables to 28 core predictors. After removing any remaining incomplete observations, the final beta regression dataset included 463 unique players. For the longitudinal analysis, each player-season was treated as a separate observation, with the season converted to a numeric variable (`SZNAB`) indicating the sequential year of a player's career.

Table 1: Description of Variables in the Dataset.

Variable	**Description**	**Variable**	**Description**
HGT	Player's standing height in inches	LPVERT	Max vertical jump in inches
WGT	Player's weight in pounds	LANE	Lane agility drill time in seconds
BMI	Body Mass Index (kg/m^2)	SPRINT	$\frac{3}{4}$ court length sprint time
WNGSPN	Wingspan in inches	MP	Total minutes played in college
STNDVERT	Max vertical jump	FG	Field goals made in college
FG%	Field goal percentage in college	FGA	Field goals attempted in college
3P	Three-pointers made in college	3PA	Three-pointers attempted in college
3P%	Three-point percentage	FT	Free throws made in college
FT%	Free throw percentage	FTA	Free throws attempted in college
ORB	Offensive rebounds	TRB	Total rebounds in college
PF	Personal fouls	AST	Assists made in college
PPG	Points per game in college	RPG	Rebounds per game in college
APG	Assists per game	PER	Player Efficiency Rating
PER_normalized	Scaled PER	POS	Primary playing position
GP	NBA games played	SZNAB	NBA season year
NPTS	NBA points per game	NAST	NBA assists per game
NREB	NBA rebounds per game	NSTL	NBA steals in a season
NBLK	NBA blocks in a season	NTOV	NBA turnovers in a season
MPG	NBA minutes per game	SZNAB	The NBA season the player played in

The distributions of the variables in Table 1 is shown in the Appendix of Figure 2 which shows a bar graph for the categorical Position variable and histograms for the continuous predictors. The variables exhibit regular distributions without extreme outliers. In this analysis, we decided to model and predict the professional player efficiency (PER) of an NBA player using the collected data from NBA statistics, college data, and NBA Combine measurements. PER is used to measure a player's performance not only in a game but throughout the season. Looking at NBA statistics both offensively and defensively, we

found that the most important statistics that decide a player's performance are points, rebounds, assists, steals, blocks, and turnovers. In our data, we had the game averaged for points, rebounds, and assists, but for steals, blocks, and turnovers, we had the total amount throughout the season. Therefore, we had to divide by the number of games they played for these values. The way the NBA calculates this formula depends on the league's averages in that particular season; however, these averages are not displayed publicly and are rather calculated internally without an exact formula being released as to how to calculate it. For this reason, we decided to create our own formula using weights to ensure that each statistic is used fairly in the formula. Firstly, to regulate the number of minutes someone would play, we used the average number of minutes played in the league (36) and divided it by the minutes per game each specific player played. Then, since points are a larger number, on average, they are multiplied by 0.8 to make them smaller. Similarly, rebounds and assists are values that are generally between 0-12, and so we decided to inflate their values to match the points by 1.5 times and 1.2 times their original value, respectively. Steals and blocks are usually smaller values around 0-5, and we decided to inflate those values by 2 and 1.7, respectively, to account for their smaller values. Lastly, we subtracted the turnovers since those hurt the player's performance as they hurt their team's chances to win; once again, this is a smaller value, causing us to double its value in the formula. To sum it up, the specific formula is:

$$\text{PER} = \frac{36}{\text{MPG}} \times \left(0.8\,\text{NPTS} + 1.5\,\text{NREB} + 1.2\,\text{NAST} + 2\frac{\text{NSTL}}{\text{GP}} + 1.7\frac{\text{NBLK}}{\text{GP}} - 2\frac{\text{NTOV}}{\text{GP}} \right).$$

Since these variables are used to compute the formula for PER, they are not regressed and used to predict the PER, and we take them out of the list of predictors for both the Beta Regression and Longitudinal model.

2 Theoretical Framework

2.1 Beta Regression Model

Suppose we observe n sets of measurements of predictor variables $x_1, \ldots, x_k$ and a continuous response variable y that takes values strictly between 0 and 1. The beta regression model, originally proposed in [8], is well-suited for modeling such responses. It assumes

that y follows a beta distribution with probability density function

$$f(y) = \frac{y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}}{B(\mu\phi,(1-\mu)\phi)}, \quad 0 < y < 1,$$

where the normalizing constant

$$B(\mu\phi,(1-\mu)\phi) = \int_0^1 y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}\,dy$$

is the beta function. The location parameter μ depends on the predictors $x_1, \ldots, x_k$ through a logistic function:

$$\mu = \frac{\exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}},$$

and the scale (or precision) parameter ϕ is a positive real number. For this distribution, the mean and variance of y are

$$\mathbb{E}(y) = \mu, \text{ and } \mathbb{V}\text{ar}(y) = \frac{\mu(1-\mu)}{1+\phi}.$$

Thus, μ represents the expected value of the response, while ϕ allows modeling the variance independently of the mean.

The parameters in this regression model are $\beta_0, \ldots, \beta_k$ and ϕ. They are estimated from data using the maximum likelihood method. The likelihood function for the dataset is the product of the individual densities:

$$L(\beta_0, \ldots, \beta_k, \phi \mid y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{y_i^{\mu_i\phi-1}(1-y_i)^{(1-\mu_i)\phi-1}}{B(\mu_i\phi,(1-\mu_i)\phi)},$$

and the corresponding log-likelihood function is

$$\ln L(\beta_0, \ldots, \beta_k, \phi) = \sum_{i=1}^{n} \Big[(\mu_i\phi - 1)\log y_i + ((1-\mu_i)\phi - 1)\log(1-y_i) - \log B(\mu_i\phi,(1-\mu_i)\phi)\Big]$$

where

$$\mu_i = \frac{\exp\{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}\}}{1 + \exp\{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}\}}.$$

Maximum likelihood estimates of the parameters are obtained by numerically maximizing this log-likelihood function. In practice, iterative optimization algorithms are used, as closed-form solutions are not available due to the complexity of the beta function. Once the regression slopes are estimated, $\widehat{\mu}$ can be computed as:

$$\widehat{\mu} = \frac{\exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k\}}{1 + \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k\}},$$

and ϕ is estimated by a numeric value $\hat{\phi}$.

2.2 Hypothesis Testing for Regression Parameters

In beta regression, inference on the regression coefficients $\beta_0, \ldots, \beta_k$ is typically performed to assess whether each predictor has a significant effect on the response. The null hypothesis for each coefficient is

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0, \quad j = 0, \ldots, k,$$

which corresponds to testing whether the associated predictor has no effect on the mean of the response variable. A common approach for testing includes the Wald test with the test statistic defined as

$$W_j = \frac{\widehat{\beta}_j^2}{\text{Var}(\widehat{\beta}_j)},$$

which is asymptotically chi-square distributed with 1 degree of freedom under H_0.

2.3 Random-Intercept Beta Regression for Longitudinal Data

Consider longitudinal measurements y_{ij} collected from n subjects ($i = 1, \ldots, n$) at J_i time points ($j = 1, \ldots, J_i$). Let $x_{ij1}, \ldots, x_{ijk}$ denote covariates associated with observation y_{ij}. To account for correlation among repeated measurements within the same subject, we introduce a subject-specific random intercept u_i. The random-intercept beta regression model is

$$y_{ij} \mid u_i \sim \text{Beta}(\mu_{ij}\phi, (1-\mu_{ij})\phi),$$

$$\mu_{ij} = \frac{\exp\{\beta_0 + \beta_1 x_{ij1} + \cdots + \beta_k x_{ijk} + \beta_{k+1} time_j + u_i\}}{1 + \exp\{\beta_0 + \beta_1 x_{ij1} + \cdots + \beta_k x_{ijk} + + \beta_{k+1} time_j + u_i\}}, \quad i = 1, \ldots, n, \quad j = 1, \ldots J_i,$$

where the random intercepts u_i's are assumed independent and normally distributed with a constant variance σ_u^2.

Conditional on u_i, the observations y_{ij} are independent, with mean and variance

$$\mathbb{E}(y_{ij} \mid u_i) = \mu_{ij}, \quad \mathbb{V}\text{ar}(y_{ij} \mid u_i) = \frac{\mu_{ij}(1-\mu_{ij})}{1+\phi}.$$

Next, the parameters of this model, $\beta_0, \ldots, \beta_k, \phi$, and σ_u^2 are estimated via a likelihood method. The conditional likelihood for subject i is

$$L_i(\boldsymbol{\beta}, \phi \mid u_i) = \prod_{j=1}^{a_i} \frac{y_{ij}^{\mu_{ij}\phi - 1}(1-y_{ij})^{(1-\mu_{ij})\phi - 1}}{B(\mu_{ij}\phi, (1-\mu_{ij})\phi)},$$

and the marginal likelihood is obtained by integrating over the random intercept u_i:

$$L_i(\beta_0, \ldots, \beta_k, \phi, \sigma_u^2) = \int_{-\infty}^{\infty} \left[\prod_{j=1}^{J_i} \frac{y_{ij}^{\mu_{ij}\phi - 1}(1 - y_{ij})^{(1-\mu_{ij})\phi - 1}}{B(\mu_{ij}\phi, (1-\mu_{ij})\phi)} \right] \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\Big(- \frac{u_i^2}{2\sigma_u^2} \Big)\, du_i.$$

The full marginal likelihood for all subjects is

$$L(\beta_0, \ldots, \beta_k, \phi, \sigma_u^2) = \prod_{i=1}^{n} L_i(\beta_0, \ldots, \beta_k, \phi, \sigma_u^2).$$

Maximum likelihood estimates of the parameters are obtained by numerically maximizing the marginal log-likelihood function. Because the integral over u_i does not have a closed form, numerical methods are typically used in practice.

2.4 Goodness-of-Fit Deviance Test

How well a model fits the data can be assessed using the deviance test (also called the asymptotic likelihood ratio test).[9] In this test, the null hypothesis is that the *null model* has a better fit, and the alternative hypothesis is that the fitted model is better. For the beta regression described in Subsection 2.1, the null model is the intercept-only model without predictors. In mathematical terms, the hypotheses are stated as follows:

$$H_0 : \ \beta_1 = \beta_2 = \cdots = \beta_k = 0 \ \text{ (intercept-only model is adequate)},$$
$$H_1 : \ \text{at least one } \beta_j \neq 0 \ \text{ (fitted model with predictors fits better)}.$$

The test statistic, called the deviance, is defined as

$$\text{deviance} = -2\Big(\ln L(\text{null model}) - \ln L(\text{fitted model}) \Big)$$

where $L(\cdot)$ denotes the maximized likelihood under a given model. Large values of the deviance indicate that the fitted model improves the likelihood substantially compared with the null model, that is, the predictors are needed.

Under H_0, the test statistic has asymptotically a chi-squared distribution, and the p-value is calculated as the upper-tail probability above the test statistic. The number of degrees of freedom is the difference between the number of parameters of the fitted and null models. The fitted model has $k + 2$ parameters ($\beta_0, \ldots, \beta_k$, and ϕ), and the null model has 2 parameters (β_0 and ϕ), thus the number of degrees of freedom is k.

For the random-intercept model introduced in Subsection 2.3, the natural test is whether the variance of the random intercept equals zero. The hypotheses are

$$H_0 : \ \sigma_u^2 = 0 \ \text{ (no random intercept)}$$
$$H_1 : \ \sigma_u^2 > 0 \ \text{ (random intercept is present and improves fit).}$$

The nominal difference in parameter count is 1 (the variance parameter). Thus, the deviance has an approximate chi-squared distribution with one degree of freedom.

3 Applications and Results

This section presents the application of the beta regression model for cross-sectional data and the random-intercept beta regression model for longitudinal data. These models are used to analyze the relationship between player characteristics and the normalized Player Efficiency Rating with multiple predictors, using the formula defined in Subsection 1.3.

3.1 Beta Regression

The beta regression model was initially fitted using all predictors. Non-significant predictors were then removed through backward elimination, resulting in a final model that includes only the significant predictors, shown in Table 2. Additionally, player position (POS) was included as a categorical variable. To facilitate interpretation, the Center (C) position was set as the reference (baseline) category. Consequently, all position coefficients represent the estimated effect relative to Centers, holding other variables constant.

Table 2: Significant Predictors in Beta Regression Model.

Predictor (Std. Units)	Estimate (log-odds)	Std. Error	z value	p-value
(Intercept)	-1.618	0.264	-6.118	< 0.001
WGT	1.250	0.605	2.065	0.039
RPG	0.170	0.054	3.156	0.002
MPG	-0.254	0.045	-5.587	< 0.001
FG	0.309	0.101	3.046	0.002
POSPG	1.231	0.352	3.495	< 0.001
POSPG-SG	1.299	0.346	3.758	< 0.001
POSSF	0.561	0.257	2.181	0.029
POSSG	0.896	0.297	3.021	0.003
POSSG-PG	0.976	0.353	2.762	0.006
POSSG-SF	0.695	0.291	2.384	0.017

The reduced fitted model with all predictors significant at the 5% level is given by

$$\text{logit}\big(\widehat{\mathbb{E}}(PER)\big) = -1.618 + 1.250 \cdot \text{WGT} + 0.170 \cdot \text{RPG} - 0.254 \cdot \text{MPG}$$

$$+ 0.309 \cdot \text{FG} + 1.231 \cdot \text{POSPG} + 1.299 \cdot \text{POSPG-SG} + +0.561 \cdot \text{POSSF}$$

$$+ 0.896 \cdot \text{POSSG} + 0.976 \cdot \text{POSSG-PG} + 0.695 \cdot \text{POSSG-SF},$$

and $\hat{\phi} = 9.64$. The p-value for the deviance test is less than 0.001, indicating a good fit.

Key results reveal the following:

- Positively Related: A one–unit increase in player weight is linked to higher log-odds of PER, suggesting that heavier players tend to have greater efficiency. Rebounds per Game (RPG) also show a positive relationship, where a one–standard deviation increase in RPG corresponds to a noticeable boost in PER, indicating that players who rebound more in college generally achieve higher efficiency. A one-SD increase in FG is associated with players with more field goals score more and have a better chance at being more efficient. Finally, positional effects reveal a clear hierarchy: guards and wing players (PG, PG-SG, SG, SG-SF) demonstrate significantly higher efficiency compared to the baseline group of centers, showing that certain positions are consistently associated with higher PER values.

- Negatively Related: A one-SD increase in MPG shows a decrease in the log-odds of PER, highlighting the more a player plays, the lower their PER will be. This shows how fatigue and health plays a crucial role in a player's success in the NBA.

Model evaluation using a 90/10 train/test split (416 training observations and 47 test observations) yielded strong predictive performance. The prediction accuracy of the model within the tolerance range of 10, 15, and 20 shows that 14.9% of predictions fall within ±10% of the observed values, 31.9% within ±15%, and 34.0% within ±20%. Figure 1 shows the predicted versus actual plot with the smoothed predicted curve generally following the trajectory of the actual values, while deviations reflect individual variability and player-specific effects in the beta regression model.

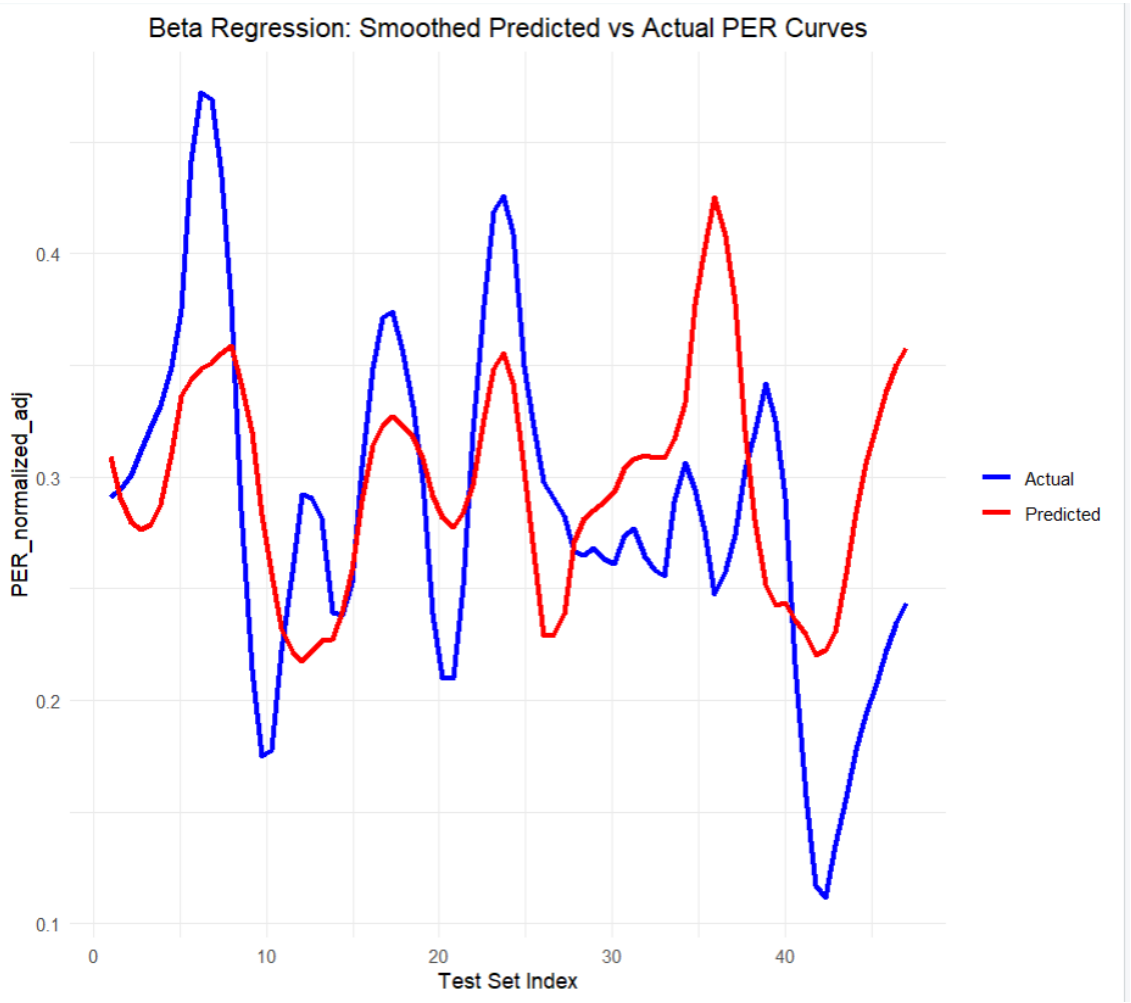

Figure 1: Predicted vs. actual normalized PER for testing set.

3.2 Random Intercept Beta Regression for Longitudinal Data

To account for the longitudinal structure of repeated player observations, a random-intercept beta regression model was fitted. This specification allows for player-specific baseline efficiency while estimating the effects of physical and performance predictors on normalized efficiency ratings. Direct components of the PER formula were excluded to avoid circularity. The set of predictors was then reduced via backward elimination to include only those significant at the 5% level. The estimated regression coefficients are presented in Table 3 that follows.

Table 3: Significant Predictors in Random-Intercept Beta Regression Model.

Predictor	Estimate	Std. Error	z value	p-value
Intercept	18.319	8.717	2.102	0.036
HGT	-0.290	0.113	-2.564	0.010
WGT	0.061	0.021	2.934	0.003
FG%	1.787	0.570	3.134	0.002
FT%	0.957	0.317	-3.015	0.003
RPG	0.040	0.015	2.638	0.008
STNDVERT	0.030	0.015	2.041	0.041
POSPG	0.927	0.243	3.810	< 0.001
POSPG-SG	1.047	0.237	4.410	< 0.001
POSSF	0.489	0.177	2.759	0.006
POSSF-PF	0.568	0.197	2.883	0.004
POSSG	0.685	0.207	3.315	< 0.001
POSSG-PG	0.811	0.244	3.327	< 0.001
POSSG-SF	0.663	0.198	3.339	< 0.001
Season (SZNAB)	0.034	0.003	12.433	< 0.001

The fitted reduced random-intercept beta regression model has the form:

$$\text{logit}\Big(\widehat{\mathbb{E}}(PER)\Big) = 18.319 - 0.290 \cdot \text{HGT} + 0.061 \cdot \text{WGT} + 1.787 \cdot \text{FG\%} + 0.957 \cdot \text{FT\%}$$
$$+ 0.040 \cdot \text{RPG} + 0.030 \cdot \text{STNDVERT} + 0.927 \cdot \text{POSPG} + 1.047 \cdot \text{POSPG-SG} + 0.489 \cdot \text{POSSF}$$
$$+ 0.568 \cdot \text{POSSF-PF} + 0.685 \cdot \text{POSSG} + 0.811 \cdot \text{POSSG-PG} + 0.663 \cdot \text{POSSG-SF} + 0.034 \cdot \text{SZNAB}.$$

The estimates of the additional parameters of this model are $\hat{\phi} = 19.178$ and $\hat{\sigma}_u^2 = 0.213$. The 95% Confidence Interval for the variance is (0.1818, 0.2490). The deviance test in this case has the p-value less than 0.05, confirming a decent fit of the model.

Interpretation of the estimated regression coefficients yields the following conclusions:

- Positively Related: Several variables show significant positive effects on PER, highlighting key factors that contribute to player efficiency. Player weight (WGT) is positively associated with PER, suggesting that heavier players, potentially due to greater strength or ability to finish around the basket, tend to achieve higher efficiency. Shooting performance is also important: both field goal percentage (FG%) and free throw percentage (FT%) are strong positive predictors, reflecting the value of scoring efficiency in overall performance. Rebounds per game (RPG) and standing vertical (STNDVERT) demonstrate that athleticism and the ability to secure possessions significantly enhance efficiency, likely due to their contributions on both offense and defense. Positional effects reveal a clear hierarchy, with guards and wings (PG, PG-SG, SG, SG-SF) consistently outperforming centers, emphasizing the greater impact of perimeter-oriented roles in modern NBA schemes. Finally, the season variable indicates that efficiency tends to improve over time, possibly reflecting player development, experience, and adaptation to professional play.

- Negatively Related: Height (HGT) shows a significant negative effect on PER, indicating that taller players generally have lower efficiency scores in this model. This may reflect the challenges that very tall players face in mobility, shooting consistency, and defensive versatility compared to guards and wings. While height can provide advantages near the basket, the negative relationship suggests that, on average, taller players may be less efficient in contributing across multiple facets of the game, particularly in perimeter-oriented or fast-paced playstyles. This underscores that physical traits alone do not guarantee high efficiency and that positional and skill-related factors play a critical role.

When evaluating predictions against actual PER values in the testing set, 15.3% of predictions fell within $\pm 10\%$ of the observed values, 21.0% within $\pm 15\%$, and 30.3% within $\pm 20\%$.

These findings align with prior research identifying college rebounding, field goal percentage, and assists as key predictors of NBA success [2, 5]. Our model extends this by capturing nonlinear effects and highlighting the relative influence of these variables. The dominance of guards and wings mirrors Mamonov's (2023) XGBoost analysis, which found perimeter players overrepresented among top performers [6]. Consistent with Moxley and Towne (2015), NBA Combine metrics such as sprint time and lane agility were not independent predictors once college performance was accounted for [4]. Overall, these results reinforce that college production is a stronger predictor of professional efficiency than physical traits alone.

4 Future Research Direction

While this study highlights the utility of beta regression and random-intercept beta regression models in predicting NBA efficiency from college and NBA Combine data, several avenues remain open for future exploration. One direction is to expand the dataset to include more recent draft classes and international players, as their inclusion could improve generalization across different styles of play. Another promising extension is to integrate advanced tracking data, such as player movement and spacing metrics, which may capture aspects of efficiency not reflected in traditional box-score statistics. Methodologically, alternative statistical approaches such as hierarchical Bayesian models or machine learning ensembles could be compared with the beta regression approach to evaluate gains in predictive performance. Together, these directions would strengthen the robustness of predictive frameworks and enhance their practical relevance for scouts, analysts, and team decision-makers.

Supplemental Materials

Both datasets used for fitting beta regression and random-intercept beta regression, along with the R codes used to run the analysis in this study, are readily available in the GitHub repository at https://github.com/areenjain09/per-prediction-regression.

Acknowledgments

A special thank you goes to Dr. Olga Korosteleva, Professor of Statistics at California State University, Long Beach, for her unwavering guidance, support, and mentorship throughout this project. Her insights and encouragement were instrumental in strengthening my understanding of statistics and machine learning, as well as in fostering my growth as a researcher. I am also very grateful to my high school statistics teacher, Ms. Machado, for not only introducing me to this opportunity but also supporting me throughout the journey. She introduced me to my passion in statistics and helped me build my foundational skills. I would like to express my heartfelt gratitude to all my teachers and mentors who have played a crucial role in shaping my academic journey and helping me become who I am today. Lastly, I extend my deepest appreciation to my family for their constant love, encouragement, and belief in me. They always made sure to be by my side throughout the process. Their support has been invaluable throughout this journey.

References

[1] Masaru Teramoto, Chad L. Cross, Robert H. Rieger, Travis G. Maak, and Stuart E. Willick. Predictive Validity of NBA Draft Combine on Future Performance. *Journal of Strength and Conditioning Research*, 32(2):396–408, 2018.

[2] Rajesh Kannan, Bryan Boudreaux, Ritwik Sengupta, and Nivedita Kannan. Predicting National Basketball Association Success: A Machine Learning Approach. *SMU Data Science Review*, 1(3), 2018.

[3] Ryan Edwards, Chris House, and Nathan Lord. Using Pre-NBA Draft Data to Project Success in the NBA. 2015.

[4] Jerad H. Moxley and Tyler J. Towne. Predicting Success in the National Basketball Association: Stability & Potential. *Psychology of Sport and Exercise*, 16:128–136, 2015.

[5] Alexander C. Greene. The Success of NBA Draft Picks: Can College Careers Predict NBA Winners? *Culminating Projects in Applied Statistics*, 1(4), 2015.

[6] Nazarii Mamonov. Maximizing Draft Outcomes: An ML-Based Approach to NBA Draftee Success, 2023.

[7] Megan Czasonis, Mark Kritzman, Cel Kulasekaran, and David Turkington. How To Predict the Performance of NBA Draft Prospects. MIT Sloan Research Paper 6955-23, MIT Sloan School of Management, September 2023.

[8] Silvia L. P. Ferrari and Francisco Cribari-Neto. Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.

[9] Samuel S. Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

Appendix

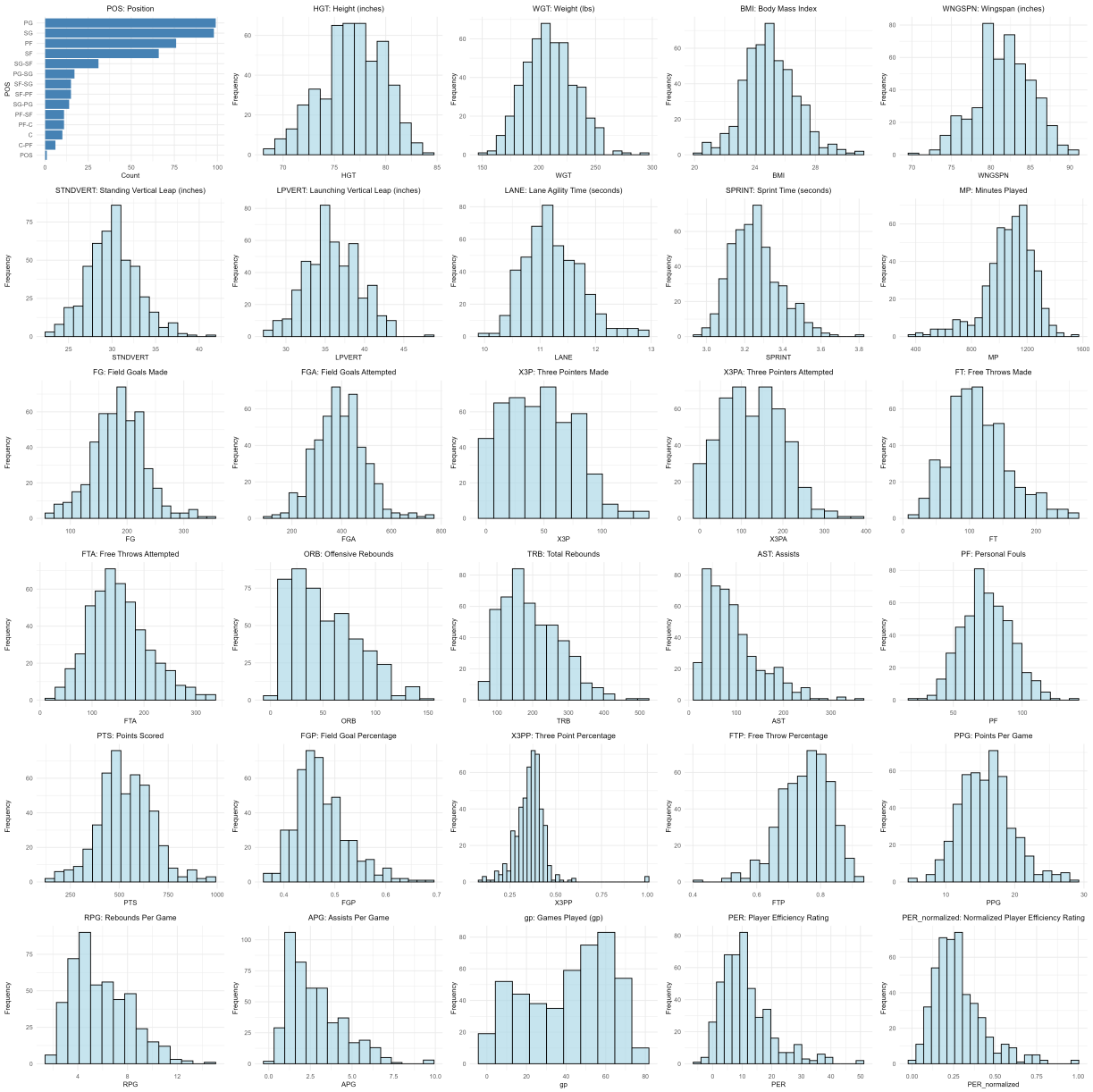

Figure 2: Bar graph for the categorical predictor and histograms for all continuous predictors in the dataset.