Performance and Pay in the NBA

Sasha Bakker, Steven Tran, Kevin Tran

Statistics is widely used throughout basketball games to predict outcomes and rank players. We plan to analyse whether or not these rankings are used in the calculations of their contract salaries.

INTRODUCTION

The question we plan to answer is “how are player salaries for the 2019-2020 NBA season correlated with their performance in the previous season?”

To answer the question, we look at 307 NBA players’ current salaries[1] next to their last season’s game data[2], taken from the basketball-reference website, to analyze whether or not there is a correlation between the two. We use performance metrics that factor in a player’s game data, when comparing performance to the contract salaries of the players (Fig.1).

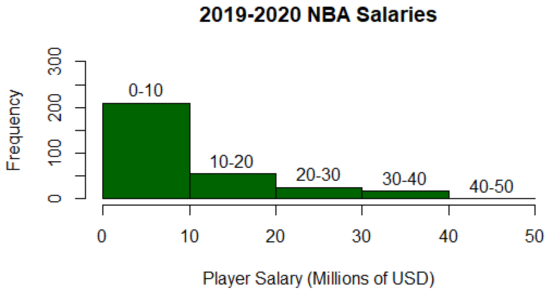


FIGURE 1. Histogram of 2019-2020 NBA player salaries in millions of U.S. dollars, showing that most salaries are under 10 million US$.

The dataset from basket-ball reference is summarized as follows:

- There are 307 valid player entries.
- Each entry corresponds to a different NBA player in the 2018-2019 season that signed one contract for the 2019-2020 season.
- The response values are the NBA player contract salaries for the 2019-2020 season.

The dataset provides comprehensive statistics about each player’s performance game by game. In order to calculate performance of each player, we wanted to use a mix of simple metrics and standard performance rank benchmarks. The chosen covariates are:

- Player Efficiency Rating (PER) - A per-minute rating of a player's performance (Fig.2a).
- NBA Efficiency (EFF) - The most commonly used player efficiency benchmark (Fig.2b).
- Effective Field Goal% (eFG%) - A player’s weighted field goal percentage (Fig.2c).

Outliers: In our data sets (Fig.3) there are numerous player entries that seem to be outliers in performance and salary. Since our metrics use percentages in calculations, these players either consistently score high or have played very well in a small number of games. The former is usually considered “NBA All-stars” and are paid much higher due to the popularity and notoriety they provide the teams they play on.

[1] https://www.basketball-reference.com/contracts/players.html
[2] https://www.basketball-reference.com/leagues/NBA_2019_per_game.html

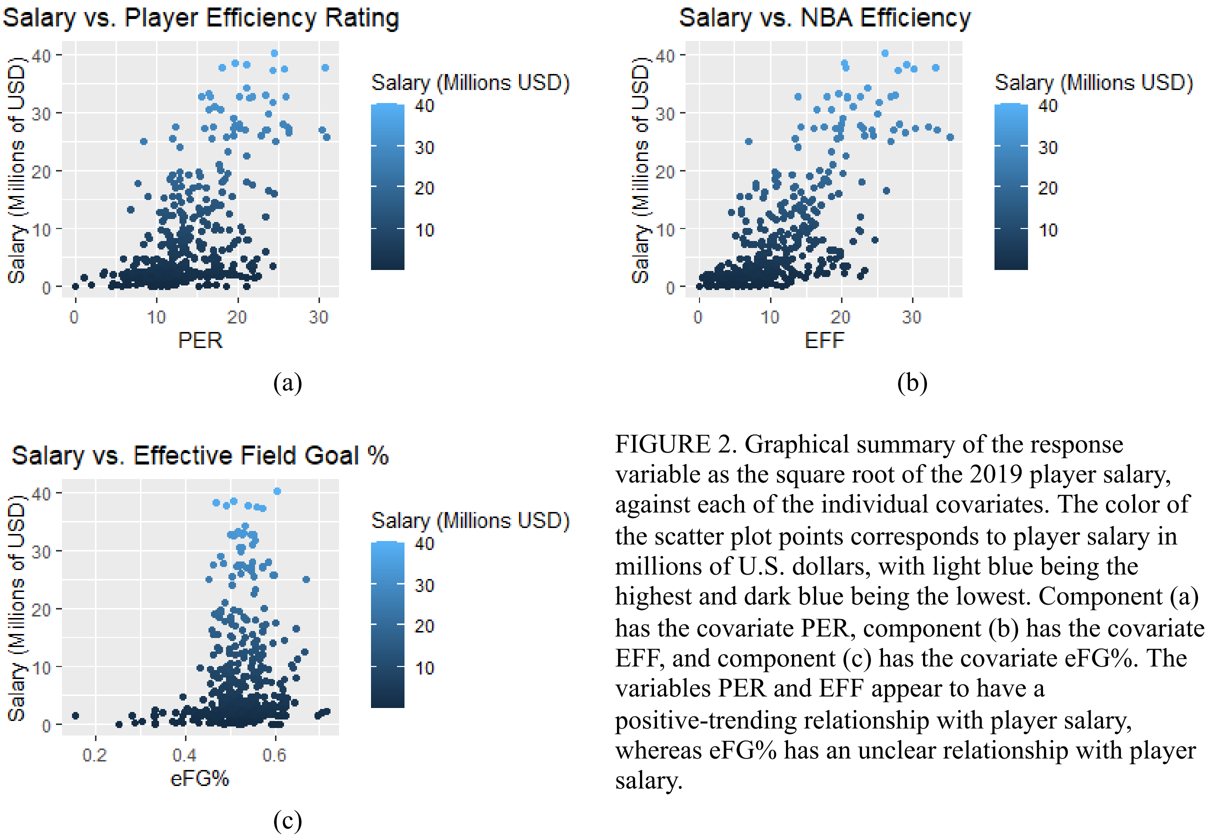


FIGURE 2. Graphical summary of the response variable as the square root of the 2019 player salary, against each of the individual covariates. The color of the scatter plot points corresponds to player salary in millions of U.S. dollars, with light blue being the highest and dark blue being the lowest. Component (a) has the covariate PER, component (b) has the covariate EFF, and component (c) has the covariate eFG%. The variables PER and EFF appear to have a positive-trending relationship with player salary, whereas eFG% has an unclear relationship with player salary.

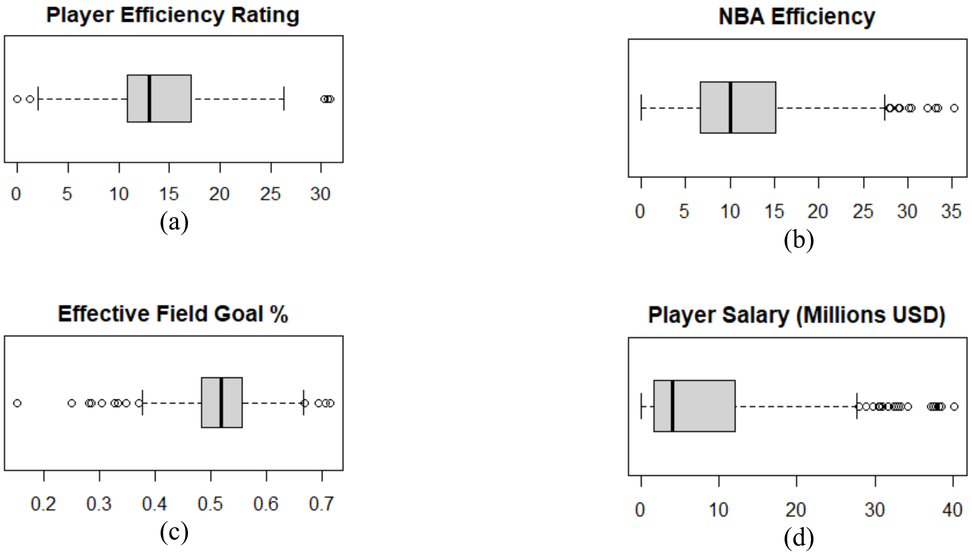


FIGURE 3. Box plots of the covariates and the response variable. (a) PER boxplot. (b) EFF boxplot. (c) eFG% boxplot. (d) 2019 salary in millions US$ boxplot.

PRELIMINARY FINDINGS

NBA Player Efficiency Rating

Player Efficiency Rating (PER) sums up all a player's positive accomplishments, subtracts the negative accomplishments, and returns a per-minute rating of a player's performance. A PER of 15 can typically indicate approximate net worth of players, and a PER of 30 or over is considered exceptionally high[3].

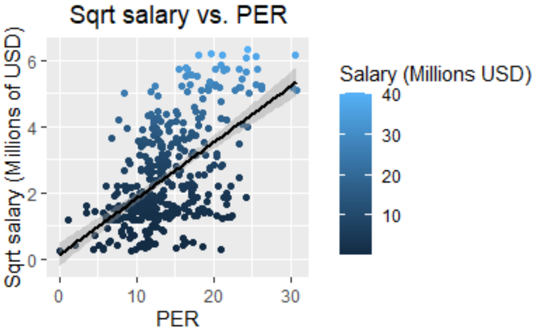
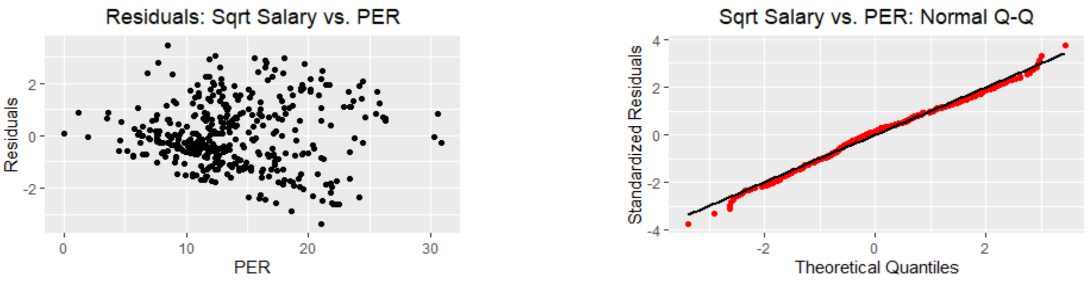


FIGURE 4. Plot of the square root of salary in the unit 'millions of USD' versus PER with a linear fit. The model for the fit is $Y = \beta_0 + \beta_1 X$ where Y is the square root of salary in the unit 'millions of USD', and X is the PER.

Slope ($\boldsymbol{\beta}_1$)	0.17± 0.01
Intercept ($\boldsymbol{\beta}_0$)	0.1 ± 0.2
R^2 Score	0.322
P-value	<2e-16

TABLE 1. Regression results for the model described in Figure 4.



(a) (b)

FIGURE 5. Assessment of residuals for the model described in Figure 4. Component (a) is a plot of the residuals and component (b) is a plot of the standardized residuals.

The square root of the response is taken to help visually linearize the data (Fig 4). The estimated slope of the square root 2019 Player Salary vs. Player Efficiency Rating is 0.17 ± 0.01 in the unit Millions of U.S. dollars. The P-value is below 2×10^{-16}, showing at the $\alpha = 0.05$ level that the slope is statistically significant. The R^2 Score of the fit is approximately 32% which suggests the model does not explain much of the variation in the square root of salary around its mean (Table 1). The residual plot of this data shows error values within the range (-3, 3), which do not appear to have constant variance across PER. However, they do appear randomly scattered about zero, which is an ideal distribution (Fig. 5a). Along with this, the Normal Q-Q plot is linear which signifies the distribution of residuals is approximately normal,

[3] https://en.wikipedia.org/wiki/Efficiency_(basketball)

thus satisfying the regression assumptions (Fig. 5b). Due to these observations, we can conclude that player performance, as measured by PER, is correlated with the player salaries in the previous season.

NBA Efficiency (EFF)

One of the major calculations the NBA performed in the past was the efficiency metric of a player. This is a simple metric, that measures the overall positives per game with the formula[4]:

$$(\text{PTS} + \text{REB} + \text{AST} + \text{STL} + \text{BLK} - \text{Missed FG} - \text{Missed FT} - \text{TO}) / \text{GP}$$

The results rank each player based on their positives they bring to a game on average, increasing the score when a player scores more, and decreasing it for each missed shot. Specific to this model, the salary is square rooted for each player because it is expected that the salary of a player grows exponentially as their exposure grows with their performance.

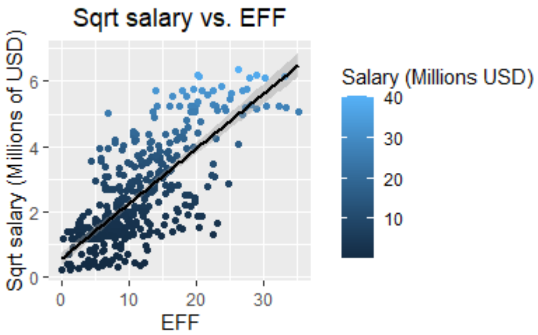
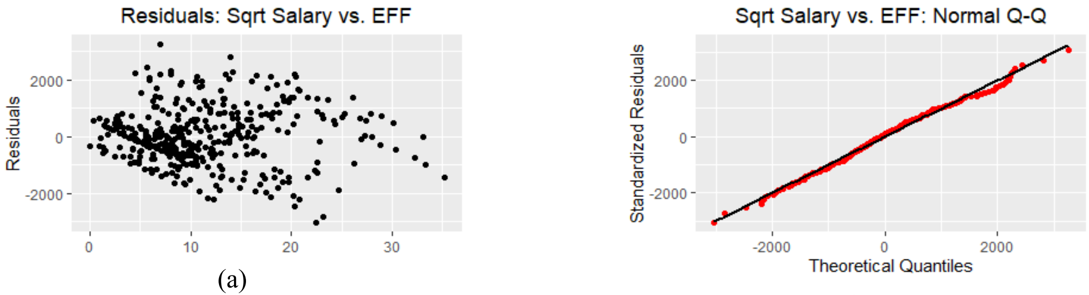


FIGURE 6. Plot of the square root of salary in the unit 'millions of USD' versus EFF with a linear fit. The model for the fit is $Y = \beta_0 + \beta_1 X$ where Y is the square root of salary in the unit 'USD', and X is the EFF.

Slope (β_1)	169 ± 8
Intercept (β_0)	563 ± 100
R^2 Score	0.5487
P-value	< 2e-16

TABLE 2. Regression results for the model described in Figure 6.



(a)

(b)

FIGURE 7. Assessment of residuals for the model described in Figure 6. Component (a) is a plot of the residuals and component (b) is a plot of the standardized residuals.

The square root of the response is taken to help visually linearize the data (Fig 6). The estimated slope of the square root 2019 Player Salary vs. NBA Efficiency is 169 ± 8 in the unit U.S. dollars. The P-value is below 2×10^{-16}, showing at the $\alpha = 0.05$ level that the slope is statistically significant. The R^2 Score of the fit is approximately 55% which suggests the model

[4] https://web.archive.org/web/20161127134724/http://www.nba.com/statistics/efficiency.html

explains some of the variation in the square root of salary around its mean (Table 2). The residual plot of this data shows error values within the range (-3000, 3000), which appears to have semi-constant variance across EFF. They appear randomly scattered about zero, which is an ideal distribution (Fig. 7a). Along with this, the Normal Q-Q plot is linear which signifies the distribution of residuals is approximately normal, thus satisfying the regression assumptions (Fig. 7b). Due to these observations, we can conclude that player performance, as measured by EFF, is correlated with the player salaries in the previous season.

NBA eFG%

The effective field goal percentage is a statistic instituted after the introduction of the three point line. This statistic takes into account three point field goals are worth more than normal field goals. It is given by[5]:

$$\text{eFG\%} = (\text{FG} + 0.5 * \text{3P})/(\text{FGA})$$

FG represents the field goals made, 3P represents the 3-point field goals made, and FGA represents the field goal attempts.

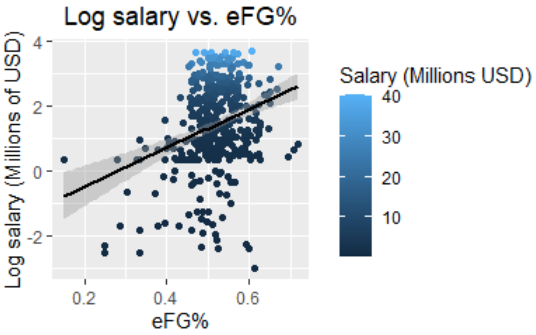
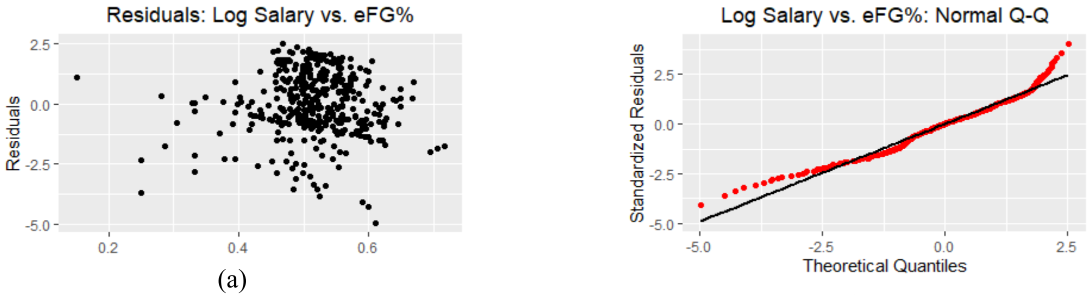


FIGURE 8. Plot of the log of salary in the unit 'millions of USD' versus eFG% with a linear fit. The model for the fit is $Y = \beta_0 + \beta_1 X$ where Y is the log of salary in the unit 'USD', and X is the eFG%.

Slope (β_1)	5±1
Intercept (β_0)	12.6 ±0.6
R^2 Score	0.07
P-value	2.2e-6

TABLE 3. Regression results for the model described in Figure 6.



(a)

(b)

[5] https://en.wikipedia.org/wiki/Effective_field_goal_percentage#:~:text=In%20basketball%2C%20effective%20field%20goal,only%20count%20for%20two%20points.

FIGURE 9. Assessment of residuals for the model described in Figure 8. Component (a) is a plot of the residuals and component (b) is a plot of the standardized residuals.

The log of the response is taken to help visually linearize the data (Fig 8). It also increases the value of R^2 from 3%, which is very low. The estimated slope of the log 2019 Player Salary vs. Effective Field Goal Percent is 5 ± 1 in the unit U.S. dollars. The P-value is 2.2×10^{-6}, showing at the $\alpha = 0.05$ level that the slope is statistically significant. The R^2 Score of the fit is approximately 7% which suggests the model explains almost no variation in the log of salary around its mean (Table 3). The residual plot of this data shows error values within the range (-5, 2.5), which appears to not have a constant variance across eFG%. They do not appear randomly scattered about zero because there are more points above zero than below zero, which is not an ideal distribution (Fig. 9a). Along with this, the Normal Q-Q plot is semi-linear which signifies the distribution of residuals is somewhat normal. This does not satisfy the regression assumptions (Fig. 7b). Due to these observations, we cannot conclude that player performance, as measured by eFG%, is correlated with the player salaries in the previous season. Further analysis must be done in order to make the result more clear.

MULTIPLE VARIABLE MODEL RESULTS AND METHODS

$$\text{Sqrt(Salary)} = Y = \beta 0+\beta 1(\text{PER})+\beta 2(\text{eFG\%})+\beta 3(\text{EFF})$$

	Coefficient	Std. Error
Intercept	1047.11	402.62
PER	-36.95	20.63
eFG%	-485.41	930.20
EFF	193.45	13.88

TABLE 4. Coefficients and Standard Errors of three covariate model

In table 4, we used a simple linear fit with PER, EFG, and EFF as covariates. This time, we notice that the coefficient for EFG has large standard error, which suggests that it may be an unnecessary variable. This is because the estimated slope is not a significant value with large uncertainty.

	eFG%	EFF	PER
P values (> | t |)	0.60	0.04e-34	0.074

TABLE 5. P-values of the multiple linear regression

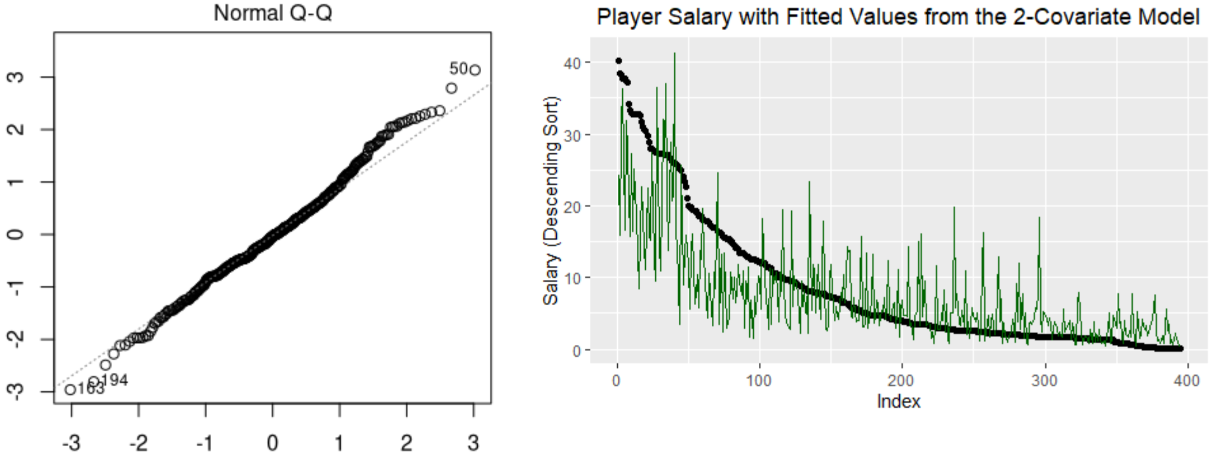
Computing the p-values of the multiple linear regression, to see whether any of these variables can be removed, at the alpha = 0.10 level, we see that PER and EFF both reject the null hypothesis that the coefficient is zero. However, EFG does not reject the null with its large p-value (Table 5). Due to this, we can remove the EFG variable from our fit.

$$\text{Sqrt(Salary)} = Y = \beta 0+\beta 1(\text{EFF})+\beta 2(\text{PER})$$

	Coefficient	Std. Error
(Intercept)	838.93	157.1
EFF	194.43	13.58
PER	-40.91	17.82

TABLE 6. Coefficients and Standard Errors of two covariate model

Referring to Table 6, both of the standard errors on the slope coefficients are fairly lower than the estimated values. Computing the coeff. Of determination R^2, we see that the fit accounts for 55% of the variance, similar to just having EFF as a covariate. Although this value is not high, it makes sense given the separate fits of the variables in Tables 1, 2, and 3. Along with this, the normal QQ plot is linear, with only some deviations near the ends of the data. This signifies that the residuals for the plot are normally distributed (Fig. 10a). However, comparing the fitted values to the actual data, there appears to be a large variance for prediction and the overall trend of the fitted values does not precisely follow the salary data (Fig. 10b).



(a)

(b)

FIGURE 10. Visualizations of the fit in Table 6. (a) Normal QQ plot of the two covariate model. (b) Comparison of the player salary with the two covariate model's fitted values for salary, based on each player's corresponding PER and EFF values.

PAIRWISE INTERACTION

From our findings in our Multivariable Regression we found that while adding our covariates together came to similar results as our Single variable regression, we found that the interaction between the covariates to be pretty high and that it is reasonably so, as the covariates are all derived from the same set of data (Fig 11).

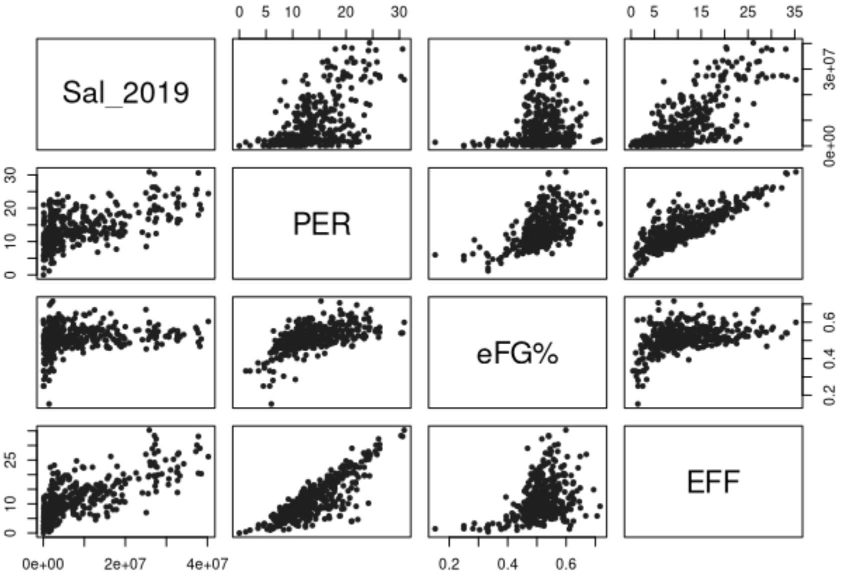


FIGURE 11. Scatterplot Matrices of our covariates and response

In Figure 11, we see that there are significant relationships between the covariates we chose for our dataset. The relationship between PER and EFF shows very strong signs of a positive linear relationship with each other, which is expected as they are both good metrics for the performance of a player. Otherwise eFG% seems to not be a very good indicator of performance of a player, as some defensive players may not try to score as many as those on offensive positions. Both PER and EFF factor that into their calculations and it looks like the best models for our data is our single variable regression between the Salary and either PER or EFF.

CONCLUSION

We found our models for PER and EFF, visualized in Table 6, to reasonably represent our data for the question. Although the R^2 values are less than 0.60, the transformed model is very linear with a significantly small p-value. Our model for eFG% did not fit the data well, shown by the low R^2 value, yet still had a low P-value. Based on our R^2 values and p-values for the PER and EFF models, there is reasonable evidence there is a correlation between player salaries and their performance in the previous year. All regression assumptions are satisfied for our PER and EFF models

NBA Player Efficiency Rating

The estimated slope of the 2019 Player Salary vs. Player Efficiency Rating is 0.170 ± 0.012 with units as the square root of Millions of USD and a P-value below 2×10^{-16}, showing at the $\alpha = 0.05$ level that the slope is statistically significant. The R^2 Score of the fit is approximately 32% which suggests the model does not explain much of the variation in the square root of salary around its mean. The residual plot of this data shows error values within the wide range (-3, 3) which do not appear to have constant variance across PER. However, they do appear randomly scattered about zero, which is an ideal distribution. Along with this, the Normal Q-Q plot is linear which signifies the distribution of residuals is approximately normal, thus satisfying the regression assumptions. Due to these observations, we can conclude that player performance, as measured by PER, is correlated with the player salaries in the previous season.

NBA Efficiency (EFF)

The resulting regression fits a more linear model with a square rooted Salary. Based on the following graphs, the residuals look randomly scattered, which suggest constant variance, whilst the normal probability plot is linear with small tails, suggesting a normal distribution. Since the regression is shown to have a very linear model when transformed, with a significantly small p-value, we can conclude that there is some relationship between a player's calculated efficiency and their salary. However, this is a weak regression still, as the R^2 value is under 60%, which may suggest that there could be a better covariate, with some signs of correlation.

NBA eFG%

The regression data fit more linearly with salaries being log transformed. Based on the residual plot, the variance is not constant because the residuals do not seem randomly distributed. The regression model has a very small linear relationship that has a slope of 5.41 with a high p-value. On top of this R^2 is 0.07 which is significantly lower than what is typically required to pull any reliable predictive estimators from the model. In this data set there were no outliers. Typically, outliers are values not contained within 1.5 standard deviations away from the means on both sides. Since the variance was so high, all values in the data set were contained.

Multivariate

From the above analyses, we can conclude that the models we analysed here are consistent with the single variable conclusions. From the coefficient of determination and the p-value of the analysis, there is reasonable evidence that there is correlation between the players' scores and their salaries, especially the p-value, which remains significantly small similar to the single variate models.

Furthermore, the residuals of the reduced model seems to be randomly scattered about zero, with the normal probability graph having small tails and a linear slope, which doesn't show any blatant concerning violations of our assumption that the model fits a normal linear relationship.

While this model looks good, there is concern for overfitting, as it doesn't really make sense for us to group together our covariates as they all derive from the same scoring statistics. Even one of our covariates we analysed, eFG%, is used in calculating both EFF and PER, which explains why we could remove it in our reduced model. For this reason, rather than analyzing the numerical interactions of our covariates, looking at the qualitative variables such as the type of salary contract a player signs will most likely get us a better sense of the relationship at hand.

R Code

Creation of the Dataset we worked on and summary of our single variable regression

```
> nba <- read_csv("Raw_data/Scores.csv")
> nba_advanced <- read_csv("Raw_data/Scores_advanced.scv")
> Salary <- read_csv("Raw_data/Salary.csv")

> eFG <- data.frame(Player = nba$Player, `eFG%` = nba$`eFG%`)
> eff <- data.frame(Player = nba$Player, EFF = ((nba$PTS + nba$TRB + nba$AST +
nba$STL + nba$BLK) - ((nba$FGA - nba$FG) + (nba$FTA - nba$FT) + nba$TOV)) / nba$G)
> per <- data.frame(Player = nba_advanced$Player, PER = nba_advanced$PER)

> eFG = eFG[!duplicated(eFG$Player),]
> eff = eff[!duplicated(eff$Player),]
> per = per[!duplicated(per$Player),]
> Salary = Salary[!duplicated(Salary$Player),]

> nba_stats = merge(eFG, Salary[,c("Player", "2019-20")], by="Player")
> nba_stats = merge(eff, nba_stats, by="Player")
> nba_stats = merge(per, nba_stats, by="Player")
> names(nba_stats)[names(nba_stats)=="2019-20"] <- "sal_2019"

> nba_stats$"sal_2019" = as.numeric(gsub("[$]", '', nba_stats$"sal_2019")) ** 0.5

> summary(lm(nba_stats$"sal_2019"~nba_stats$`eFG%`))
> summary(lm(nba_stats$"sal_2019"~nba_stats$EFF))
> summary(lm(nba_stats$"sal_2019"~nba_stats$PER))
> pairs(nba_stats[,-1], pch=20, col="grey12")
```

NBA Player Efficiency Rating (PER) Single-Variable Model (Figures 4 and 5, Table 1)

```
> library(tidyverse)
> load("C:/Users/asus/Documents/Stat_525/IE/nba_stats. RData")
> nba <- data.frame(nba_stats)
> nba$sal_2019 <- nba$sal_2019 / 1e6

> x <- nba$PER
> y <- sqrt(nba$sal_2019)

> linmod <- lm(y-x)
> e <- linmod$residuals
> bo <- linmod$coefficients["(Intercept)"]
> bi <- linmod$coefficients["x"]

> summary(linmod)

> n <- length(x)
> MSE <- sum(12)/(n - 2)
> qs <- (1:n - 0.5)/(n)
> Ee <- qnorm(qs, mean = 0, sd = sqrt(MSE))
> nba$Ee <- Ee
> nba$e <- e

> ggplot (data = nba, aes(x = PER, y = sqrt(sal_2019), col= sal_2019)) +
      geom_point() +
      xlab("PER") +
      ylab("Sqrt salary (Millions of USD)") +
      ggtitle("Sqrt salary vs. PER") +
      theme(plot.title = element_text(hjust = 0.5)) +
      geom_smooth(method='lm', formula=y~x, col="black")
> ggplot (data = nba, aes(X = PER, y = e)) +
      geom_point() +
      xlab("PER") +
      ylab("Residuals") +
      ggtitle("Residuals: sqrt salary vs. PER") +
      theme (plot.title = element_text(hjust = 0.5))
> ggplot (data = nba, aes(x = sort(e), y = Ee)) +
      geom_point(col="red") +
      geom_smooth(method='1m', formula= y~x, col="black") +
      ylab("Standardized Residuals") +
      xlab("Theoretical Quantiles") +
      ggtitle("Sqrt Salary vs. PER: Normal Q-Q") +
      theme(plot.title = element_text(hjust = 0.5))
```

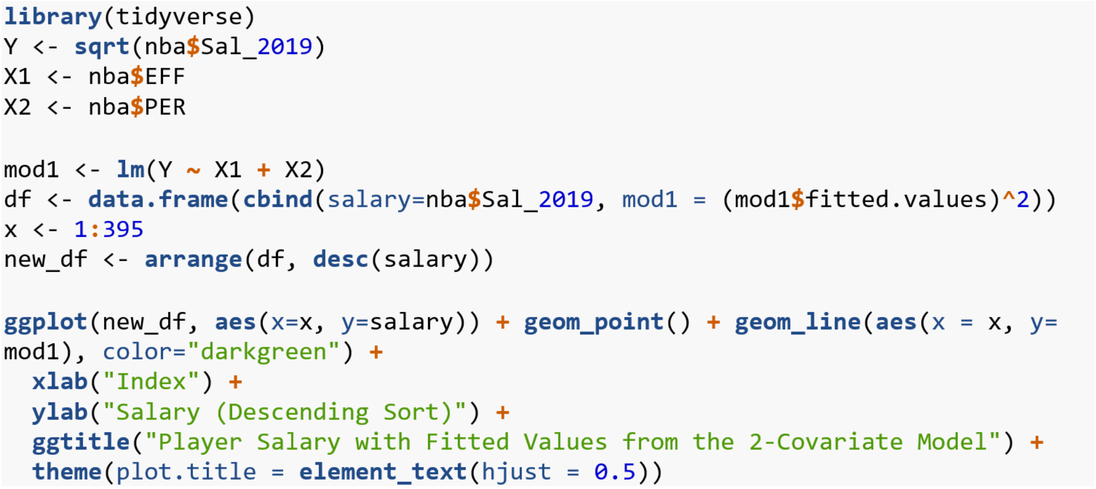
NBA Efficiency (EFF) Single-Variable Model (Figures 6 and 7, Table 2)

```
> library(tidyverse)
> load("C:/Users/asus/Documents/Stat_525/IE/nba_stats. RData")
> nba <- data.frame(nba_stats)

> x <- nba$EFF
> y <- sqrt(nba$sal_2019)

> linmod <- lm(y-x)
> e <- linmod$residuals
> bo <- linmod$coefficients["(Intercept)"]
> bi <- linmod$coefficients["x"]
> summary(linmod)

> n <- length(x)
> MSE <- sum(12)/(n - 2)
> qs <- (1:n - 0.5)/(n)
> Ee <- qnorm(qs, mean = 0, sd = sqrt(MSE))
> nba$Ee <- Ee
> nba$e <- e

> ggplot (data = nba, aes(x = EFF, y = sqrt(sal_2019), col= sal_2019)) +
      geom_point() +
      xlab("EFF") +
      ylab(Sqrt salary (Millions of USD)") +
      ggtitle("Sqrt salary vs. EFF") +
      theme(plot.title = element_text(hjust = 0.5)) +
      geom_smooth(method='lm', formula=y~x, col="black")
> ggplot (data = nba, aes(X = EFF, y = e)) +
      geom_point() +
      xlab("EFF") +
      ylab("Residuals") +
      ggtitle("Residuals: sqrt salary vs. EFF") +
      theme (plot.title = element_text(hjust = 0.5))
> ggplot (data = nba, aes(x = sort(e), y = Ee)) +
      geom_point(col="red") +
      geom_smooth(method='1m', formula= y~x, col="black") +
      ylab("Standardized Residuals") +
      xlab("Theoretical Quantiles") +
      ggtitle("Sqrt Salary vs. EFF: Normal Q-Q") +
      theme(plot.title = element_text(hjust = 0.5))
```

NBA Effective Field Goal Percent (eFG%) Single-Variable Model (Figures 8 and 9, Table 3)

```
> library(tidyverse)
> load("C:/Users/asus/Documents/Stat_525/IE/nba_stats. RData")
> nba <- data.frame(nba_stats)

> x <- nba$eFG%
> y <- log(nba$sal_2019)

> linmod <- lm(y-x)
> e <- linmod$residuals
> bo <- linmod$coefficients["(Intercept)"]
> bi <- linmod$coefficients["x"]

> summary(linmod)
> n <- length(x)
> MSE <- sum(12)/(n - 2)
> qs <- (1:n - 0.5)/(n)
> Ee <- qnorm(qs, mean = 0, sd = sqrt(MSE))
> nba$Ee <- Ee
> nba$e <- e

> ggplot (data = nba, aes(x = eFG%, y = log(sal_2019), col= sal_2019)) +
      geom_point() +
      xlab("eFG%") +
      ylab(Log salary (Millions of USD)") +
      ggtitle("Log salary vs. eFG%") +
      theme(plot.title = element_text(hjust = 0.5)) +
      geom_smooth(method='lm', formula=y~x, col="black")
> ggplot (data = nba, aes(X = eFG%, y = e)) +
      geom_point() +
      xlab("eFG%") +
      ylab("Residuals") +
      ggtitle("Residuals: log salary vs. eFG%") +
      theme (plot.title = element_text(hjust = 0.5))
> ggplot (data = nba, aes(x = sort(e), y = Ee)) +
      geom_point(col="red") +
      geom_smooth(method='1m', formula= y~x, col="black") +
      ylab("Standardized Residuals") +
      xlab("Theoretical Quantiles") +
      ggtitle("Log Salary vs. eFG%: Normal Q-Q") +
      theme(plot.title = element_text(hjust = 0.5))
```

Plot (Fig. 10b)



```
library(tidyverse)
Y <- sqrt(nba$Sal_2019)
X1 <- nba$EFF
X2 <- nba$PER

mod1 <- lm(Y ~ X1 + X2)
df <- data.frame(cbind(salary=nba$Sal_2019, mod1 = (mod1$fitted.values)^2))
x <- 1:395
new_df <- arrange(df, desc(salary))

ggplot(new_df, aes(x=x, y=salary)) + geom_point() + geom_line(aes(x = x, y=
mod1), color="darkgreen") +
  xlab("Index") +
  ylab("Salary (Descending Sort)") +
  ggtitle("Player Salary with Fitted Values from the 2-Covariate Model") +
  theme(plot.title = element_text(hjust = 0.5))
```