Does a Basketball Player in his prime Play More Than Others?

Group 9: Shakila Cham, Vinny Yabor and Thomas Connolly

December 1, 2020

Introduction and Methods

In this project we attempt to estimate the age in which a professional male basketball player reaches the prime of his career. Throughout this project we are only concerned with players within the National Basketball Association (NBA) whom were active members of the NBA in the 2019-2020 regular season. An active member of NBA refers to a player under contract who had playing time on the court sometime during the season. Regular season refers to the portion of the season before the playoff tournament. Thus we will refer to such players simply as *players.* The goal of this project is to estimate which players in the future regular season we expect to see an improvement in their minutes and which ones we expect to experience a downfall based on their age. Since we desire to make mathematical based prediction about future seasons we used data on players from the 2019-2020 regular season. Data collected from the NBA's website(NBA, 2020).

In this project we have chosen to quantify the quality of a player by the amount of time they actually play in a game. In particular we have utilized the mean minutes played per game (MPG) throughout the season by each player. Since we are making inferences about the mean MPG of a player based on their age we have divided all players into six distinct categories based on their age. The NBA had exactly 529 players in the 2019-2020 season, from ages of 19 to 43 years old (as of the beginning of the 2019-2020 season). Let $\mathcal{N}$ denote the total sample size, i.e., the total number of active players within NBA in the 2019-2020 season: $\mathcal{N} = 529$. These age categories were chosen based on age and for convenience based on the number of samples in each group being similar. The goal of similar sample sizes was not to achieve a balanced design, but to reduce the likelihood of having outliers.

First, in order to determine if a player in his prime plays more than others, we conducted a one-way ANOVA test on mean MPG by age group. Then we conducted a Scheffé post hoc test to find out which groups were different. This was followed by performing pooled sample t-tests on the unequal means to find out which age groups had more or less MPG. Next, to come up with a predictor equation for future NBA season, we performed multiple regression to determine the effect of points per game (PTS), field goals made (FGM), assists per game (AST), and steals (STL) on MPG. After coming up with the regression equation, we used the best subsets method to see if we could come up with a better predictor equation.

.

Enumerate these six categories by $C_1, C_2, C_3, ..., C_6$ and for each i, $i = 1, 2, ..., 6$ let n_i denote the size of each category i. These six categories along with their respective sizes are as follows.

		Age Groups of Players in Years					
Categories	C_i	≤ 21	22-23	24-25	26-27	28-31	32$\leq$
Enumeration	i	1	2	3	4	5	6
Category size	n_i	59	112	100	93	105	60

ANOVA Test

In this section we are going to address if age influences how much a player gets to play in a game. To do so we shall test if there is a significant difference between the mean MPG's of the 6 age categories defined above. For each category C_i, $i = 1, 2, ..., 6$ define μ_i to be the population mean of the average MPG of each category i. Further for each i, let $\widehat{\mu_i}$ be the respective sample mean of category C_i which are *unbiased* estimates of the true population mean of each of the categories. Therefore the hypotheses to be tested are

$$H_0 : \mu_1 = \mu_2 = ... = \mu_6 \qquad \text{vs.} \qquad H_a : \mu_i \neq \mu_j \text{ for some } i, j = 1, 2, ..., 6 \tag{1}$$

Here the alternative hypotheses, H_a, is *two-sided* because any variation in any direction from the equality of the population means of each category is critical. Further our sample statistics (in minutes) are given by

		Statistics of 2019-2020 Season (in minutes)					
	C_i	C_1	C_2	C_3	C_4	C_5	C_6
Sample Mean	$\widehat{\mu_i}$	18.902	16.205	17.809	20.822	23.271	20.162
Sample Variance	${s_i}^2$	84.193	96.211	84.898	81.816	61.256	66.388

The hypothesis (1) was tested by creating a one-way ANOVA table followed by interpreting the p-value. In order to conduct a one-way ANOVA, a few assumptions must be made. First, the population must follow a normal distribution. However since the sample size n_i from each category C_i, $i = 1, 2, ..., 6$ is greater than thirty, we can apply the Central Limit Theorem and assume normality. Second, the data must be independent. Each player is unique to his own group; and since the group ages are independent of each other the data is independent.

Further there must be homogeneity of variances among the categories from the population, i.e., ${\sigma_1}^2 = {\sigma_2}^2 = ... = {\sigma_6}^2$. We utilized Bartlett's test to figure if ${\sigma_1}^2 = {\sigma_2}^2 = ... = {\sigma_6}^2$ holds. The hypothesis of this test are as follows

$$\mathcal{H}_0 : {\sigma_1}^2 = {\sigma_2}^2 = ... = {\sigma_6}^2 \qquad \text{vs.} \qquad \mathcal{H}_a : \sigma_i^2 \neq \sigma_j^2 \text{ for some } i, j = 1, 2, ..., 6 \tag{2}$$

The test-statistic of this test is given by

$$T = \frac{(\mathcal{N} - k)\ln{(t_p{}^2)} - \sum_{i=1}^{6}(n_i - 1)\ln{(s_i{}^2)}}{(1 + \frac{1}{3}\cdot(k-1))\cdot((\sum_{i=1}^{6}\frac{1}{(n_i-1)}) - \frac{1}{(N-k)})}$$

In the T formula given above k is the number of variances being tested, s_i^2 is sample variance of category i, $i = 1, 2, ..., 6$ and the value $t_p{}^2$ is given by

$$t_p{}^2 = \frac{\sum_{i=1}^{6}(n_i - 1)\cdot s_i{}^2}{\mathcal{N} - k}$$

The test statistic T follows a chi-squared distribution with $1 - \alpha$ and $k - 1$ degrees of freedom where α is the level of significance. For $\alpha = 0.05$ the resulting p-value is given by $p - value = 0.2074$. This p-value is greater than significance level of $\alpha = 0.05$, so we fail to reject the null hypothesis $\mathcal{H}_0$ given in (2). Thus, we assume homogeneity of variances and henceforth proceed with conducting a one-way ANOVA test.

To proceed with this test several calculations were done. These calculations involve finding the treatment sum of squares (SSA), sum square of error (SSE), mean square of treatment (MSA), mean square of error (MSE), F-statistic, total sum of squares (SST), and degrees of freedom. The equation of these values are given by the ANOVA table as follows

Source of variation	Sum of squares	Degree of freedom	Mean Square	F
Treatments	$SSA = \sum_{i=1}^{6} n_i(\widehat{\mu}_i - \bar{\bar{\mu}})^2$	$a - 1$	$MSA = \frac{SSA}{a-1}$	$F = \frac{MSA}{MSE}$
Error	$SSE = \sum_{i=1}^{6}\sum_{j=1}^{n_i}(y_{ij} - \widehat{\mu}_i)^2$	$\mathcal{N} - a$	$MSE = \frac{SSE}{\mathcal{N}-a}$	Pr(>F)
Total	$SST = SSA + SSE$	$\mathcal{N} - 1$		

where a is the number of categories, i.e., $a = 6$, y_{ij} is the mean MPG of players j, $j = 1, 2, ..., n_i$, from category i, $i = 1, 2, ..., 6$. Furthermore, $\bar{\bar{\mu}}$ is the grand sample mean. The F-statistic follows an f distribution with $k - 1$, $\mathcal{N} - k$, α degrees of freedom. Using R and formulas above the ANOVA table is calculated to be

Source of variation	Sum of squares (SS)	Degree of freedom	Mean Square (MS)	F
Age group	3241	5	648.3	8.152
Error	41587	523	79.5	Pr($> F$)
Total	44828	528		2.01×10^{-7}

From the table, our p-value= $2.01 \times 10^{-7} \approx 0$ is much smaller than $\alpha = 0.05$, so we can reject the null hypothesis and conclude that there is at least one pair of age groups which are significantly different in terms of minutes played per game. Our next step was to determine which age groups were different. There are a few post-hoc tests we could employ such as a Tukey test, Bonferroni method, and the Scheffé method. The Tukey test is better to use when you have a balanced design. Although our sample sizes were similar, they were not exactly the same, so we used the Scheffé method. The relevant formulas are

$$t_{ij} = \frac{\widehat{\mu_i} - \widehat{\mu_j}}{\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}} \qquad C_\alpha = \{t_{ij} : |t_{ij}| > \sqrt{(k-1)F_{k-1,\mathcal{N}-k,\alpha,U}}\}$$

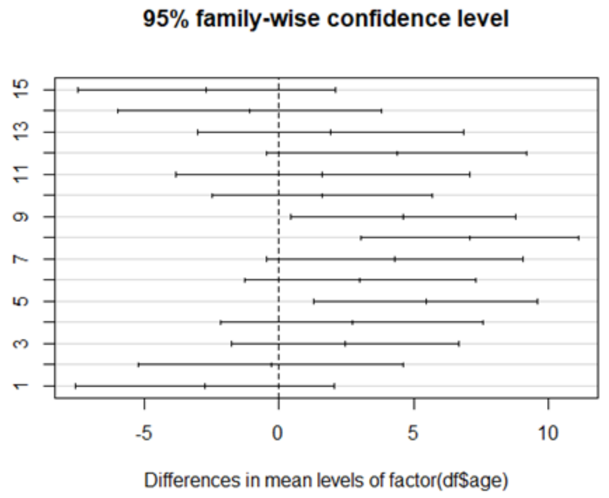
The $100 \cdot (1-\alpha)\%$ confidence interval is given by

$$(\widehat{\mu_i} - \widehat{\mu_j} \mp \sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})} \cdot \sqrt{(k-1)F_{k-1,\mathcal{N}-k,\alpha,U}})$$

After conducting the post-hoc test, we found only a few significantly small p-values. The p-values and confidence intervals are summarized in the table below:

Categories	Lower Limit	Upper limit	p-value
C_1 vs. C_2	-7.4885108	2.094049	0.6183
C_1 vs. C_3	-5.9821570	3.796767	0.9898
C_1 vs. C_4	-3.0374626	6.877083	0.8920
C_1 vs. C_5	-0.4763376	9.215805	0.1083
C_1 vs. C_6	-3.8458798	7.075823	0.9644
C_2 vs. C_3	-2.4932330	5.702304	0.8873
C_2 vs. C_4	0.4385934	8.795489	0.0192
C_2 vs. C_5	3.0210681	11.112860	3.6e-06
C_2 vs. C_6	-0.4528547	9.077260	0.1058
C_3 vs. C_4	-1.2781705	7.303181	0.3594
C_3 vs. C_5	1.3007282	9.624129	0.0020
C_3 vs. C_6	-2.1561018	7.571435	0.6301
C_4 vs. C_5	-1.7912401	6.691087	0.5904
C_4 vs. C_6	-5.2367714	4.627094	1.0000
C_5 vs. C_6	-7.5749082	2.065384	0.6021

The p-values less than 0.05 are for the age groups 26 to 27 vs 22 to 23, 28 to 31 vs 22 to 23, 28 to 31 vs 24 to 25. Thus, we conclude that players in these age groups do not play the same average MPG. Another way to interpret this test is by looking at the confidence intervals. The graph below summarizes the confidence intervals and clearly displays which ones are significant.



On the y-axis, 1 represents ages 32 and over vs 28 to 31, 2 represents ages 32 and over vs 26 to 27, and so on. These are all in reverse order from the table of p-values listed above. As you can see, 0 does not fall within the confidence intervals for the fifth, eighth, and ninth age groups. This matches up with the interpretation of the p-values in that we reject the null hypothesis and come to the same conclusion. According to a paper published by Rob Simmons and David J. Berri, NBA players typically reach their maximum performance, or prime, when they are 26 years old (Simmons and Berri 2011, 386). Our studies indicate that players who are at or slightly past their prime play a different amount of minutes per game compared to players who are less experienced. Now we test to see if players who are near or past their prime (age groups 26 to 27 and 28 to 31) play less or more minutes than the younger players (age groups 22 to 23 and 24 to 25). Since players in their prime are more efficient and more likely to contribute to wins, we speculate that they get more minutes on average. This is tested using a t-test for each of the three significantly different pairs. We are using a pooled variance t-test based on the earlier findings from the Bartlett test that the groups all share the same variance. To conduct a pooled variance t-test, one would need to compute

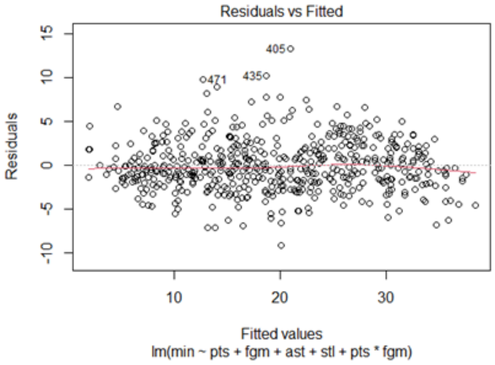
$$t_{ij} = \frac{\widehat{\mu_i} - \widehat{\mu_j}}{t_p\sqrt{(\frac{1}{n_i} + \frac{1}{n_j})}}$$

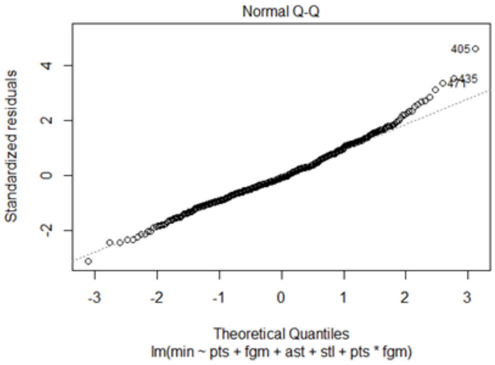
which follows a Student's t-distribution under the null. The results are the following. For ages 26 to 27 vs 22 to 23, the p-value is 0.0003121. So we reject the null and conclude that players ages 26 to 27 play more than the younger age group. For ages 28 to 31 vs 22 to 23, the p-value is 9.451e-09. So we reject the null and conclude that players ages 28 to 31 play more than the younger group. Finally, for ages 28 to 31 vs 24 to 25, the p-value is 4.013e-06. So we reject the null and conclude that players ages 28 to 31 play more than the 24 to 25 age group. Thus, we can say that players at or past their prime get more minutes on average than less experienced players.

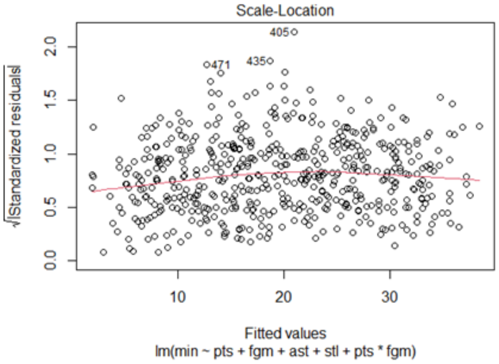
Multiple Regression Analysis

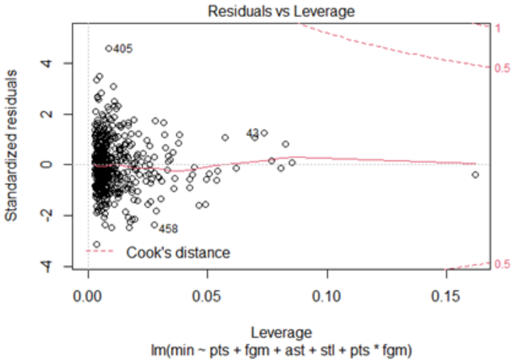
For our second test we chose to use the chapter 11 method of multiple linear regression. Multiple linear regression allows you to check the relationship between several predictor variables and a response variable. Multiple linear regression results in a model of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ In this equation y is the response variable, $x_1, ..., x_k$ are the predictor variables, and $\beta_0, ..., \beta_k$ are the known parameters that in our case need to be calculated. For our test we used minutes played as the response variable and tested various in game statistics for the predictor variables to see the result these factors would have on the number of minutes an NBA player would be put into the game for. For our predictor variables we chose points per game, assists per game, steals per game, field goals made per game, and lastly a statistic produced by multiplying the field goals made by the number of points scored per game. So, based on our multiple regression model we will be able to predict how many minutes an NBA player should theoretically receive given a certain stat line.

There are several assumptions that need to be checked to analyze the goodness of fit of a multiple regression model on a set of data. We used R to graphically check these assumptions. We made a residual vs fitted value graph. Our graph clearly shows that the assumption of constant variance holds being how close the red line is to the dotted line through the middle. We originally did not satisfy this condition which is why we added the x_2 variable representing the product of the points and the field goals made. This transformation satisfied the assumption. We created a normal plot of the residuals. The graph shows a relatively straight line meaning that the normal assumption also holds. We also checked the results of the scale-location plot to see if the residuals are equally spread. From our graph it is clear that the residuals are randomly spread and follow no sort of pattern. We then looked at the residual vs leverage graph to check for any outliers of influence among our data. In our graph Cook's distance line, the red dashed line, is almost not visible and all observations are within them, thus it is safe to say there are no influential outliers. The respective graphs are shown below:









After seeing that all of our assumptions were satisfied, we looked at the summary of the multiple regression model and interpreted the results. First you can look at the coefficients for our predictor variables as well as the intercept value.

Intercept	$pts = x_1$	$fgm = x_2$	$ast = x_3$	$stl = x_4$	$pts : fgm = {x_5}^2$
1.95	1.622	1.947	0.212	4.6678	-0.145

From which our regression model can be written out as:

$$min = 1.95 + 1.622x_1 + 1.947x_2 + 0.212x_3 + 4.6678x_4 - 0.145x_5^2$$

The next things seen in the summary is the standard error column which shows the standard error of each coefficient estimate calculated. The next column is the t-value of each coefficient which can be found by calculating $t_j = \frac{\beta_j}{SE(\beta_j)}$. Where SE is the standard error of the coefficient estimate. The next piece of significance is the p-value of each coefficient, telling us how significant each variable is in our model. The residual standard error is the standard deviation of the residuals. It is calculated by computing $s = \sqrt{\frac{\sum_{i=1}^{n} {\epsilon_i}^2}{n-(k+1)}}$. Here, n is the sample size and k is the number of groups. The multiple and adjusted R-squared values tell us the percentage of the variances of the response variables that are explained by our model. The multiple R-squared value is given by $r^2 = 1 - \frac{SSE}{SST}$ and the adjusted R-squared value is given by $r^2 = 1 - \frac{MSE}{MST}$. The F-statistic allows us to see whether at least one of our predictors is significant and can be calculated using $F = \frac{MSR}{MSE}$. The following table summarizes our regression hypotheses:

Hypothesis 1 :$H_{1,0}$: The pts variable is significant vs. $H_{1,a}$: The pts variable is not significant
Hypothesis 2 :$H_{2,0}$: The fgm variable is significant vs. $H_{2,a}$: The fgm variable is not significant
Hypothesis 3 :$H_{3,0}$: The ast variable is significant vs. $H_{3,a}$: The ast variable is not significant
Hypothesis 4 :$H_{4,0}$: The stl variable is significant vs. $H_{4,a}$: The stl variable is not significant
Hypothesis 5 :$H_{5,0}$: The $pts \cdot fgm$ variable is significant vs. $H_{5,a}$: The $pts \cdot fgm$ variable is not significant

The corresponding p-values are given by

Hypotheses i	P-value
Hypothesis 1	$< 2 \times 10^{-16} \approx 0$
Hypothesis 2	$2.56 \times 10^{-7} \approx 0$
Hypothesis 3	0.0629
Hypothesis 4	$< 2 \times 10^{-16} \approx 0$
Hypothesis 5	$< 2 \times 10^{-16} \approx 0$

From the p-values, we reject all null hypotheses with the exception of hypothesis 3 at significance level of $\alpha = 0.05$. Therefore, the only variable which is not significant on minutes per game is assists. Furthermore, the regression equation is

$$y = 1.950073 + 1.621930 \cdot x_1 + 1.946996 \cdot x_2 + 0.212012 \cdot x_3 + 4.667762 \cdot x_4 - 0.145352 \cdot x_5^2$$

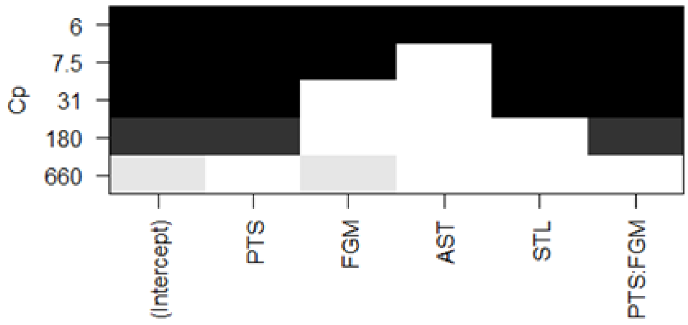
This equation could possibly be improved upon as a predictor for minutes per game. The best subsets regression method is a good way of optimizing our equation. This algorithm essentially chooses a subset of variables which optimizes a certain criterion. There are several possible criteria we could choose. These include t_p^2, adjusted t_p^2, MSE_p, C_p, and $PRESS_p$. We opted to employ the C_p criterion since it can judge well the predictive power of a model. The goal of this method is to compute and minimize the function

$$\Gamma_p = \frac{1}{\sigma^2} \cdot (\sum_{i=1}^{n} [E(\widehat{y_{ip}}) - E(y_i)]^2 + [\sum_{i=1}^{n} var(\widehat{y_{ip}})])$$

where $\widehat{y_{ip}}$ is the predicted value for Y_i, and σ^2 is the variance of residuals. Since this equation depends on unknown parameters, we may estimate it by using Mallows' C_p statistic. This is

$$C_p = \frac{SSE_p}{\sigma^2} + 2(p+1) - n$$

where p is the model size. The following graph shows several possibilities for our regression equation and how they relate to the C_p statistic.



Based on this plot, The C_p value is minimized when all five variables are included in our regression equation. Thus, no change in our equation is necessary. We already have the best predictor for minutes per game.

Effects of Missing Values

Missing values in a data set is a potential problem one may run into when performing data analysis. In this section, we investigate how missing values impact our results. There are two scenarios of interest. The first scenario is where twenty percent of our data is missing at random and in the second scenario we look at non-ignorable missing values.

Random Missing Values

For this section we simply use R to sample 423 of the 529 (%80) of the NBA players and ran our regression once more. Again, the assumptions did not hold without the interaction variable of $pts \cdot fgm$. The plots which show that the assumptions hold are shown below

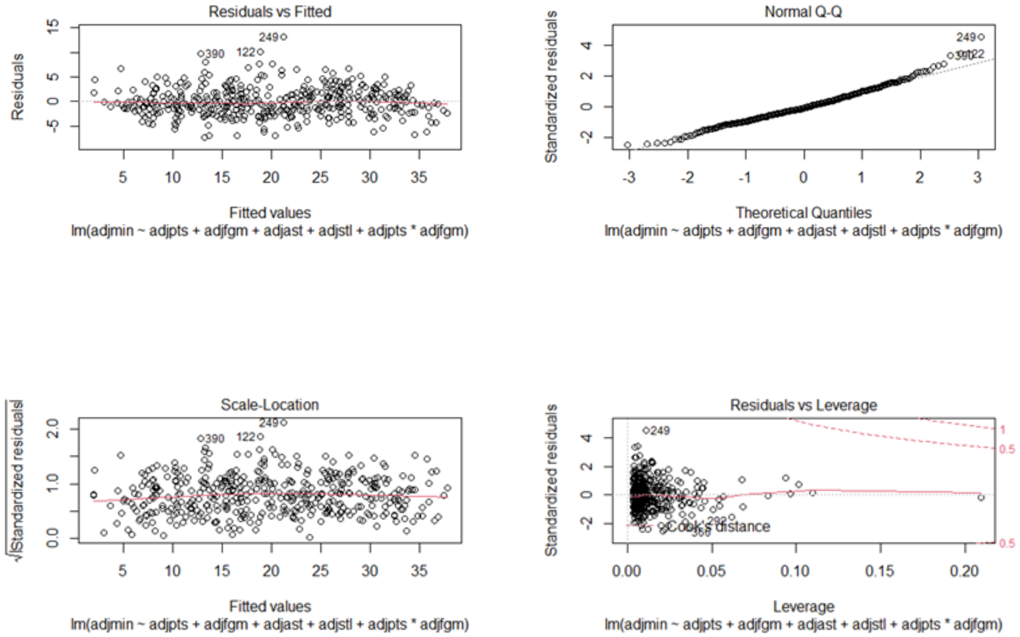
After checking the summary of the linear model function in R, we found the following results.

Intercept	$pts = x_1$	$fgm = x_2$	$ast = x_3$	$stl = x_4$	$pts : fgm = {x_5}^2$
1.95098	1.53553	2.14603	0.25399	4.92516	-0.14644

This yields a regression equation of

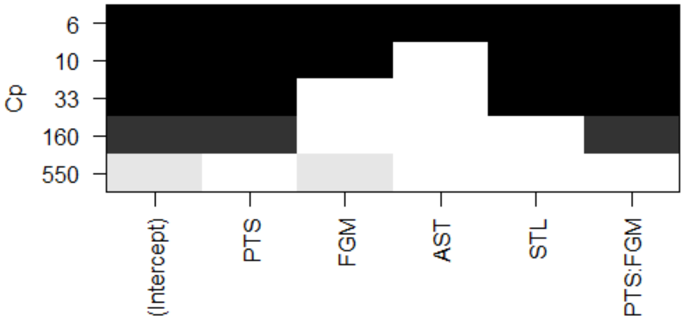
$$y = 1.95098 + 1.53553x_1 + 2.146036 \cdot x_2 + 0.25399 \cdot x_3 + 4.92516 \cdot x_4 - 0.14644 \cdot x_5^2$$

The hypotheses mirror those of our previous regression model in that we have checked the p-value for each variable to see if there is a significant interaction between each variable and the response.



Hypotheses i	P-value
Hypothesis 1	$< 2 \times 10^{-16} \approx 0$
Hypothesis 2	$2.56 \times 10^{-7} \approx 0$
Hypothesis 3	0.0442
Hypothesis 4	$< 2 \times 10^{-16} \approx 0$
Hypothesis 5	$< 2 \times 10^{-16} \approx 0$

This time, all predictor variables are significant at the $\alpha = 0.05$ level. Next, we checked the C_p criterion for the best subsets regression method. We got the same results as last time in that we already have the optimal equation for prediction.



Non-ignorable Missing Values

Professional athletes within the NBA experience a high rate of game-related injuries. When a player gets injured and still plays, he is typically put on a 'minutes restriction.' Since some players do not get put on a restriction when injured, we will limit our non-ignorable data to those who are on a minutes restriction. A minutes restriction is when a player who is playing through an injury gets less minutes each game in order to prevent them from further aggravating the injury. For example, during the 2019-2020 season, Kemba Walker who averaged 31.1 minutes per game was put on a minutes restriction of nine minutes per game after suffering pain in his left knee (Schuster, 2020).

During the 2019-2020 regular season it was reported that 349 players were injured and as a result missed between 1 and 66 games (NBA 2019 Injured Reserve Tracker). Each team in the NBA played between 63 to 75 regular season games. Although not every player who experiences injury is put on minute restriction. Given there are 349 players out of 529 players in the 2019-2020 season, the results can be significant and thus non-ignorable. A technique for handling non-ignorable missing data is to maximize the data collection (Kang, 2013). To eliminate the bias caused by injury on our data, we can eliminate those players who have been put on minute restriction out of our data set. Furthermore since this reduces our sample size we can expand our data by including data from players who were not put on minute restriction from sufficiently more regular seasons. The observed bias is caused by each player on minute restrictions which would lower the mean MPG of their respective age group, thus potentially skewing the data. Therefore eliminating these players eliminates the bias.

References

Kang, H. (2013). The prevention and handling of the missing data. Korean Journal of Anesthesiology, 64(5), 402. doi:10.4097/kjae.2013.64.5.402

NBA 2019 Injured Reserve Tracker. (n.d.). Retrieved December 01, 2020, from https://www.spotrac.com/nba/injured-reserve/2019/

NBA Traditional Regular Season Stats. (n.d.). Retrieved December 01, 2020, from https://www.nba.com/stats/players/traditional/?sort=PTS

Schuster, B. (2020, July 27). Celtics News: Kemba Walker Will Have Minutes Restriction at Restart Amid Injury. Retrieved November 30, 2020, from https://bleacherreport.com/articles/2901825-celtics-news-kemba-walker-will-have-minutes-restriction-at-restart-amid-injury

Simmons, R., and Berri, D. J. (2011). Mixing the princes and the paupers: Pay and performance in the National Basketball Association. Labour Economics, 18(3), 381-388. doi:10.1016/j.labeco.2010.11.012