Jose Mendoza MAT022 Coursework 2020-2021

21015647

09 February 2021

Abstract

This is a coursework report corresponding to the module MAT022 during the year 2020 - 2021

Contents

Introduction

The following analysis is an attempt to aproach the provided dataset using as many different methods taught in the module Foundations of Statistical Analysis and Data Science (MAT022). Even if this may not be the best approach to study this set of data when in a proffesional environment, the scope of this work is to make use of the tools learnt during the course.

Due to the limitations in time and space, the report will try to obtain significance of the results obtained and make generalizations, in detriment of an strict approach of this data.

The libraries used for this work and required to run the R-Markdown document are - readr, dplyr, tidyr, car, ggplot2, ggpubr and modelr. rstudioapi is used to set working directory to source file location.

Background

As this dataset as been approached without any knowledge of basketball, most of the assumptions and considerations may be called naive from a sports proffesional's point of view.

The approach that was made when dealing with the data had the intention to cover the datset as a whole, rather at looking at individual parameters relationship, whose importance we may don't understand. For this reason, the report focusses on the player's analysis and frequently we group and summarise data to obtain player's stats (players accuracy, average shot distance, etc.). An example of this can be found in the section (Player's Accuracy). The values from this "fabricated"" dataset are statistcs and not parameters as the data has been summarised. However we will later use these values as "attributes," in a non-strict way, to make the data "understandable" for somebody without basketball knowledge.

Lastly, need to mention that a part of this work is conditional. The R-Markdown document performs sampling during the execution of the code. For this reason, the output of some of the tests are conditional (i.e. based on resulted p-values) and the results from the report will differ when compared with individual R-Markdown file executions.

1. Whole dataset exploration

Initially we look at the dataset focussing on the most important parameter of every data entry, the result (whether the shot was made or missed).

Bernouilli experiment describing a shot

The primary objective of basketball is shooting a ball through the defender's hoop. The result will be either positive or negative (made or missed) and therefore can be considered as a Bernouilli experiment, where p is the probability of making a shot.

We will use a frequentist approach to infere the probability of success, from a sample of 100 shots.

Let $X_1, \ldots, X_n$ be the random sample shots from the Bernoulli(p) distribution, as:

$$\sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p).$$

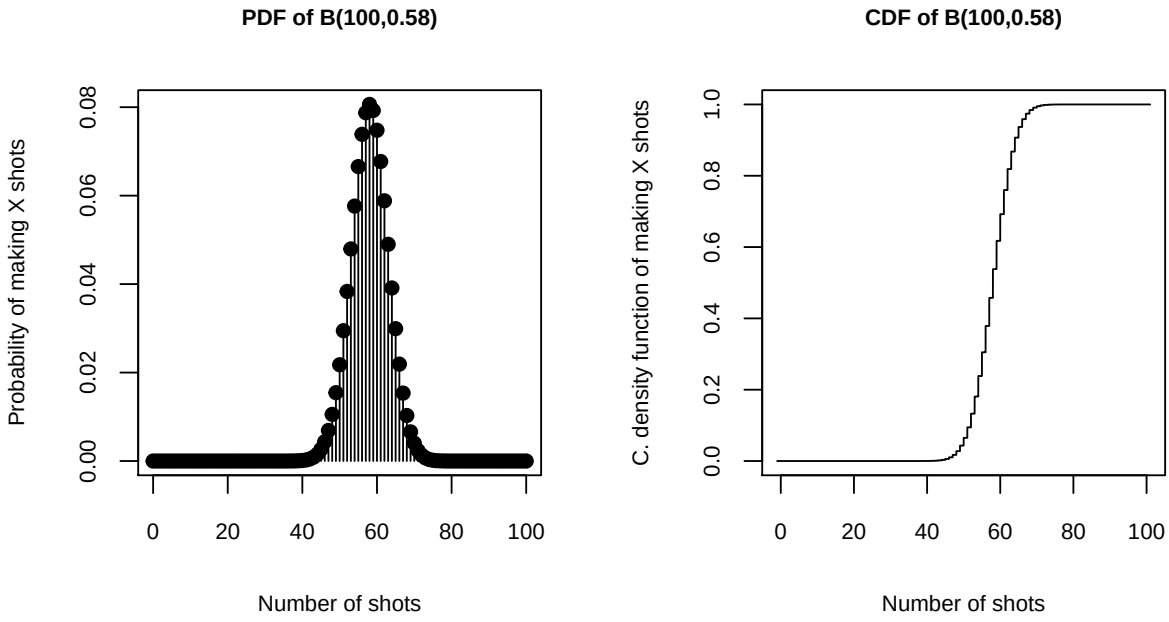
Where p is the empirical probability of making a shot obtained from a frequentist approach:

```
nba_sample_1 <- nba_df[sample(nrow(nba_df), 100), ]

all_made_1 <- nrow(nba_sample_1 %>% filter(SHOT_RESULT == 'made'))
all_missed_1 <- nrow(nba_sample_1 %>% filter(SHOT_RESULT == 'missed'))
all_shots_1 <- nrow(nba_sample_1)

p_made_1 <- all_made_1 / all_shots_1
```

The sampled probability of making a shot is:

```
## [1] 58 %
```

This can be represented as a binomial distribution, where we can obtain the probality of making X number of shots:



We create a confidence interval for this sample using the normal approximation, as the number of samples is sufficiently large.

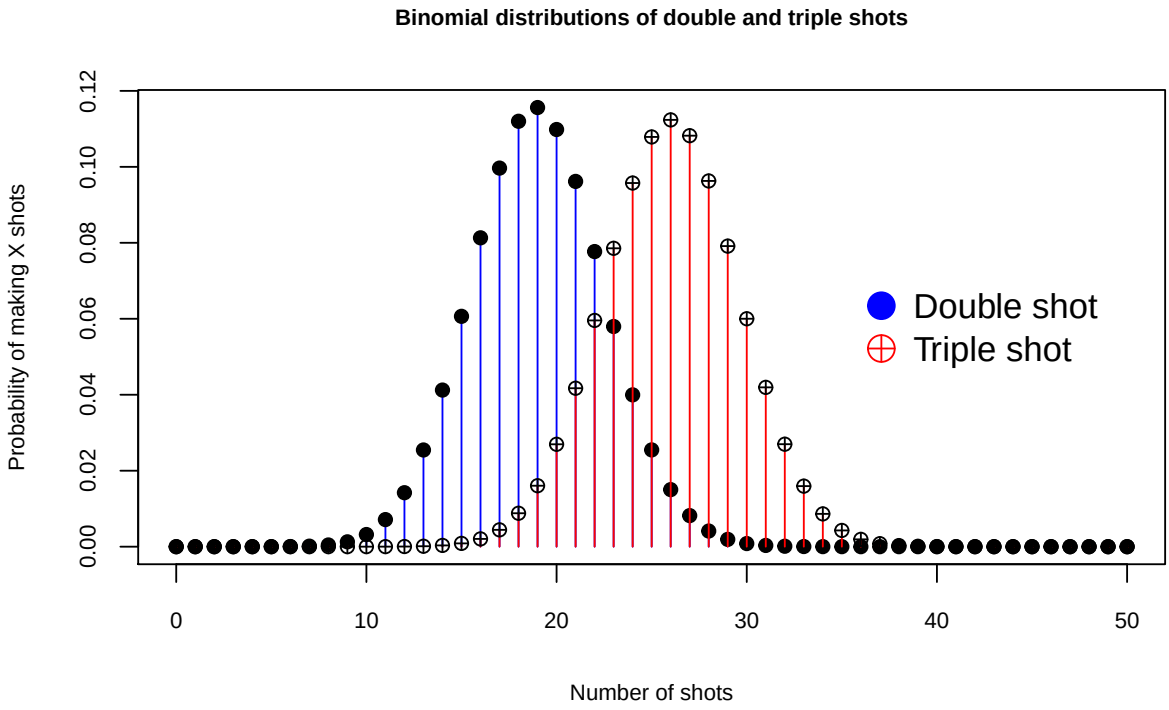
The confidence intervals:

```
##     Lower     Upper
## 0.4832643 0.6767357
```

Double vs Triple shot - equality of proportions

We are repeating the above experiment, but this time we are taking two differentiaded samples: Double and Triple shots. We calculate the binomial distributions for each sample and perform an equality of proportions test to see whether this fact affects the probability of making a shot, depending on whether the samples are found to be statistically different.

```
## [1] Equality of proportions test
```



Performing a 2-sample test for equality of proportions:

```
## [1] Number of samples (double and triple)    =
## [2] 50
## [3] 50

## [1] Empirical probabilities (double and triple)    =
## [2] 0.38
## [3] 0.52

## [1] Observations = 17              25
```

The obtained p-value and confidence intervals:

```
## [1] Lower confidence interval limit:  -0.370912967843323
```

```
## [1] Higher confidence interval limit:  0.0509129678433234
```

```
## [1] P-Value:  0.156111492960049
```

```
## [1] The probability of making a shot is the same.
```

As the difference between the two probabilities lies within the confidence region and the p-value is larger than 0.05, we retain the null hypothesis that the difference seen in the distribution graph is due to sampling error.

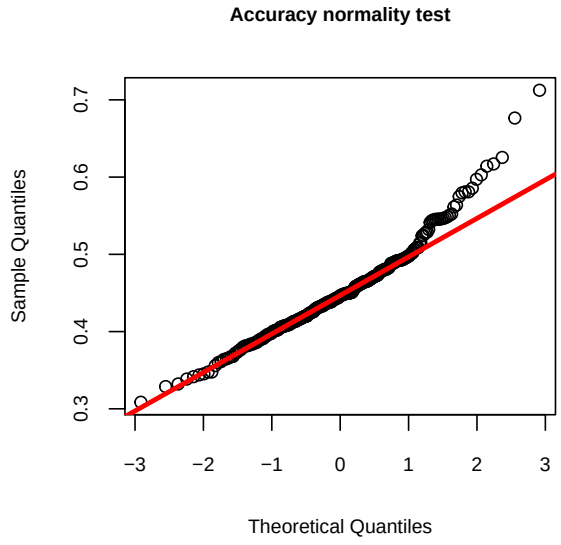
2. Player's analysis

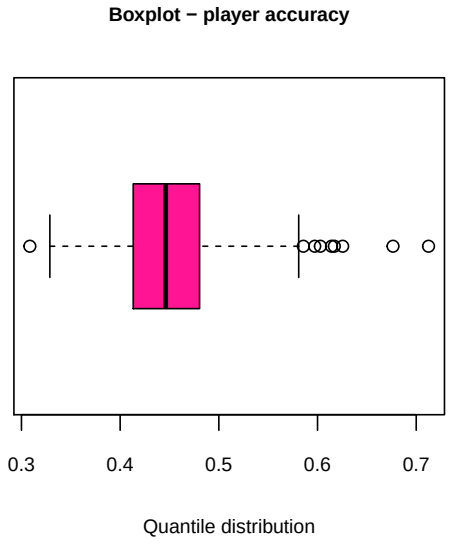
Players accuracy: Empirical probabily of making a shot

The following test may be controversial but could be a way to "label" or "rank" players, and giving an indication of player performance for the dataset.

We will compute the empirical probability of making a shot for each player, using a frequentist approach, and will call this statistic his "accuracy." As we are using the whole dataset to obtain this value, we will make the assumption that this is the true probality, to leave aside the sampling error considerations discussed above.

We will use the accuracy of the players as a ranking attribute that will be later used for more inferential analysis.






```
##  PLAYER_NAME            made           missed
##  Length:281         Min.   : 29.0   Min.   : 18.0
##  Class :character   1st Qu.:120.0   1st Qu.:142.0
##  Mode  :character   Median :191.0   Median :223.0
```

```
##                     Mean   :206.1   Mean   :249.7
##                     3rd Qu.:273.0   3rd Qu.:334.0
##                     Max.   :480.0   Max.   :580.0
##     accuracy
##  Min.   :0.3085
##  1st Qu.:0.4132
##  Median :0.4461
##  Mean   :0.4515
##  3rd Qu.:0.4805
##  Max.   :0.7125
```

In terms of the qq-plot, we see that the player's accuracy is only normally distributed on the lower and central quartiles, what reveals that in the dataset there are some players "outperforming" above the rest.

We can confirm this hypothesis when we look at the boxplot graph, where the outliers above the distribution break. We therefore conclude that the actual distribution is "right tailed" to include the outperforming players.

The quartiles of the distribution, and interquartile range:

```
## [1] "1st quartile = 0.413"
```

```
## [1] "3rd quartile = 0.480"
```

```
## [1] "IQR = 0.067"
```

Exact binomial test to study playing Home/Away effect

Earlier we obtained the accuracy for each player, which we are using a way of measuring the probability of making a shot. During the following test, we will perform an experiment in which we will select a player, and obtain its accuracy.

In this test, the null Hypothesis is that the player will make the same percentage of shots playing home or away.

```
## [1] The player randomly selected is:  Trevor Ariza
```

```
## [1] His accuracy during the 2014 season was: 38 %
```

```
## [1] From a random sample of  100  shots while playing away, he made:  38 shots.
```

```
## After performing an Exact Binomial Test, the p-value obtained was
## 0.550711824558793
```

```
##  and therefore we
## RETAIN
##  the Null Hypothesis that the performs the same playing home or away.
```

Comparison between players using a t-test

One interesting metric for a player is the amount of made shots he will perform during the season. If you are a coach and are interesting in getting a new player for the next season, you would be interested in this metric.

During the following test, we will select two random players and will calculate the number of made shots per game. We will perform a t-test under a confidence region of 95% to investigate whether there is a significant difference between both players. The Null Hypothesis is that there is no difference between the two players.

```
## [1] The two randomly selected players are:  Al Jefferson and Channing Frye
```

```
## [1] Al Jefferson  has got an average of  7.64  points per game
```

```
## [1] Channing Frye  has got an average of  2.77586206896552  points per game
```

```
## [1] The p-value obtained from the t-test is:  7.19276275064592e-17
```

```
## [1] We REJECT the null hypothesis that both players have a similar performance
```

3. Inferential analysis using player's quantitative metrics

ANOVA test to investigate effect on Game Period

The Null Hypothesis is that the Game Period does not affect in the shot clock.

```
## [1] After performing a Lavene test, the p-value was :  0.319219099236995
```

```
## [1] therefore
```

```
## [1] "the groups have a similar variance."
```

```
## [1] After performing a Shapiro test, the p-value was : 0.064855522225387
```

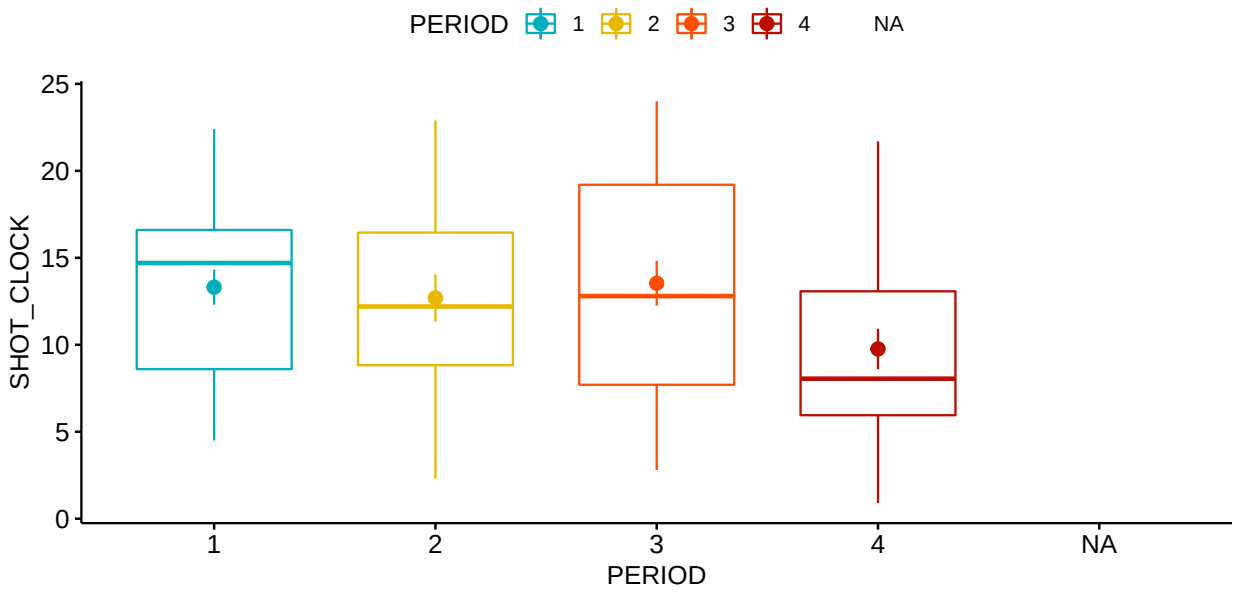
```
## [1] therefore
```

```
## [1] "The residuals are normally distributed."
```

We retain the Null Hypothesis when the p-value obtained from the Anova analysis is larger than 0.05:

```
## From the Anova analysis, the p-value was : 0.216060066837243
```

```
## [1] "the Game Period does not have an effect in the shot clock."
```



Correlation between number of dribbles and shot clock

We would like to study a numerical parameter and it's correlation with another one. Let's explore the relationship between dribbles and shot clock. One would expect that a player's performance would be conditioned with the amount of time he's got available before shooting.

We are now exploring whether there could be a relationship between two parameters: The average number of dribbles and the average shot clock, for each player, as a reference statistic of performance. For each player, we would like to see whether these parameters are correlated using the Pearson's product moment correlation. Our sample will contain one entry per player (mean statistic) and therefore considers the whole dataset.

During the calculation, we will replace the missing values with the mean of the remaining sample.

```
## [1] We compare the Average number of dribbles

## [1] and the Average shot clock for each player

## [1] The result obtained is:  -0.0953912983695966
```

As the correlation parameter is close to zero, we conclude that these two parameters are not correlated and therefore have little interest in inferential analysis.

Correlation between accuracy and shot distance

```
## [1] We compare each player's accuracy with

## [1] his corresponding average shot distance.
```

```
## [1] The result obtained is:  -0.488006280310689
```

As the obtained result differs from zero, we conclude that these parameters are sufficiently correlated to attempt to build a regression model. The negative results indicates that the less the distance the more the accuracy, as we may intuitively expect.

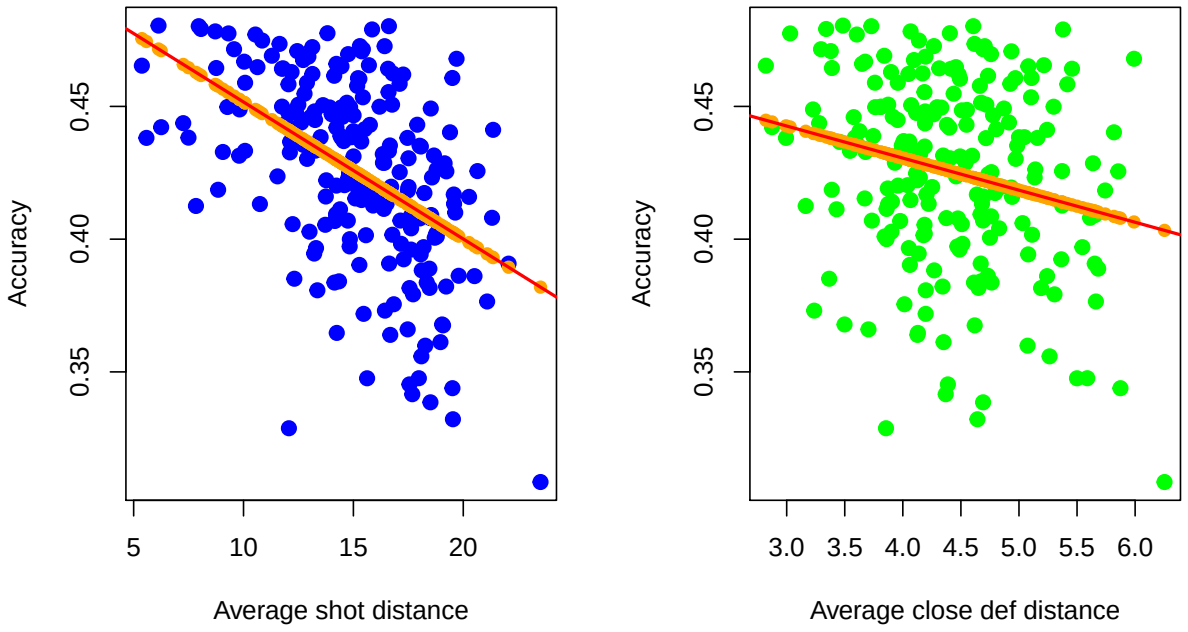
As discussed on Test 2, to build the model we are only selecting players below the third quartile to ensure we are working with a normally distributed population.

Prediction model based and average shot distance

We build two lineal regression models: ## asd_model_1 compares the average shot distance with his corresponding player's accuracy ## asd_model_2 compares the average closest defender distance with his corresponding player's accuracy

```
## [1] asd_model_1
```

```
## [1] asd_model_2
```



Model 1 analysis

```
## [1] RSE Model 1:
```

```
## [1] 0.03110261
```

```
## [1] RSE Model 1 vs mean accuracy:

## [1] 0.07310875

## [1] R-squared model 1:  0.238150129622675
```

Model 2 analysis

```
## [1] RSE Model 2:

## [1] 0.03470771

## [1] RSE Model 2 vs mean accuracy:

## [1] 0.08158277

## [1] R-squared model 2:  0.0513027496430108
```

The R-squared value indicates de variation in accuracy explained by either average shot distance or average closest defender distance.

As the R-squared value is higher in model 1, we retain this model. The variation in the response parameter is more influenced (explained) by the input variable in this model.

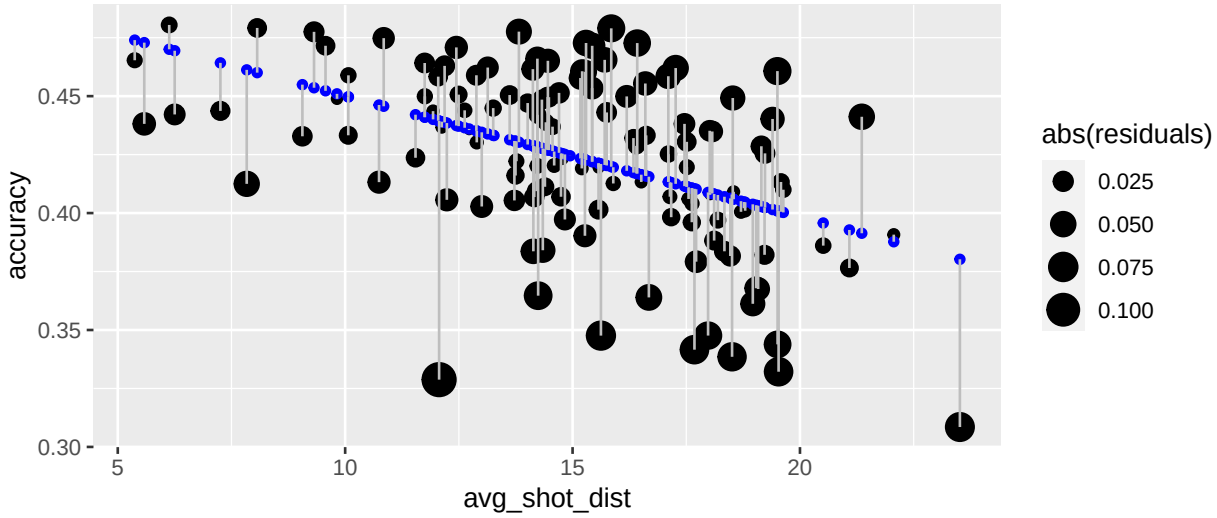
Visualize the residuals

As the residuals are the difference between the value of an outcome variable predicted by the model and the actual observed value of the variable, we need to make predictions and then testing to measure the offset. To test our model, we re-build it but this time splitting the dataset between train and test subsets. A common practice is to use 60% for train then 40% for test so we will use this criteria.

```
## [1] We retrain the model splitting the data 60-40 for train and test

## [1] The average RSE:  0.0761812229722879

## [1] The R-Squared value:  0.239613816719661
```

We are now checking the model residuals using the prediction structure. The geom_segment ggplot feature stresses the value of the points depending of the magnitud of the offset. So the most offender deviants are highlighted in the graph.

```
train$estimate <- predict(model_asd_12)
train$residuals <- residuals(model_asd_12)

plot0 <- players_regression_dataset %>%
  ggplot(aes(avg_shot_dist, accuracy)) + geom_point()

#plot0

plot <- train %>%
  ggplot(aes(avg_shot_dist, accuracy)) + geom_point(aes(size = abs(residuals)) )+ g

plot
```



Interpretation

Here we are making the assumption that the shot distance is the only variable that influences one player's accuracy. As discussed during the introduction, we are making this assumption for didactical purposes and this doesn't reflect the actual basketball gameplay.

We are looking at the intercept coefficient, which is only interpretable if we can reasonably expect a zero value for all independent variables in a model. A "zero" distance in basketball is a difficult concept to analyze as the gameplay is completely different and there are other variables with heavier influence that would impact the accuracy, and we are not considering those in this model.

```
intercept_coefficient <- model_asd_12$coefficients[2]
```

According to the simple regression analysis of Player Accuracy by Average Shot Distance, we estimate that for every additional foot, the player accuracy decreases by:

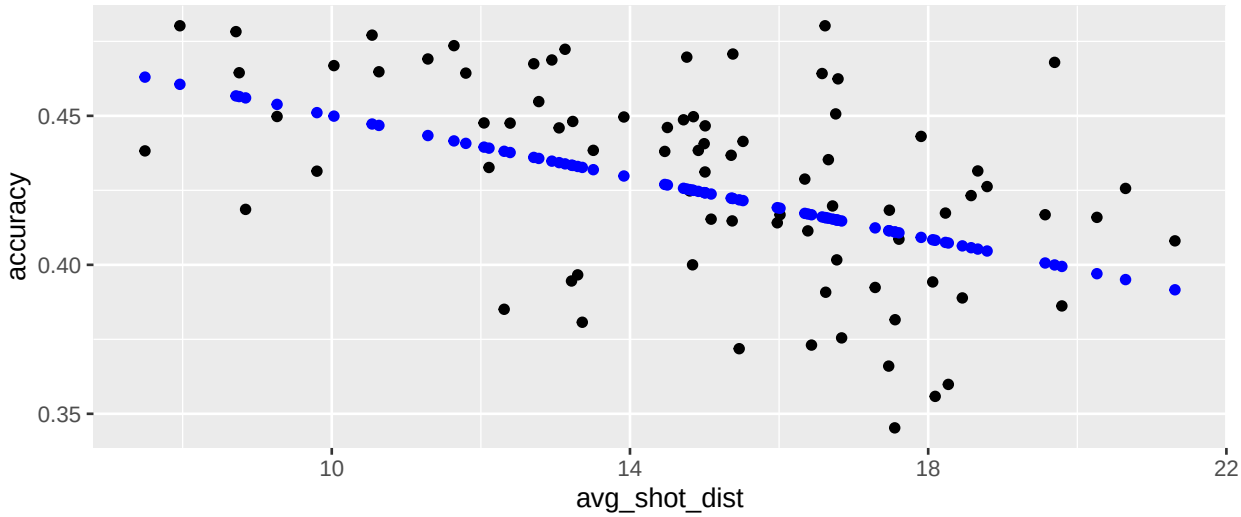
```
intercept_coefficient
```

```
## avg_shot_dist
##  -0.005166267
```

MSE and predictions

As the MSE measures the average squared difference between predicted and observed values, we are calculating this value using a summarise dplyr function feeding the split dataset "test" (40% sampled rows from the original player's accuracy dataset). This MSE will be plotted to verify to visualize the prediction.

```
mse <- test %>%
  add_predictions(model_asd_12) %>%
  summarise(MSE = mean((accuracy-pred)^2))

plot_3 <- test %>%
  add_predictions(model_asd_12) %>%
  ggplot(aes(avg_shot_dist, accuracy)) + geom_point() + geom_point(aes(y=pred), col

plot_3
```



Conclusions

The main conclusions from this analysis are that the majority of players have a similar performance with the exception of some outstanding players whose accurary is above the rest.

From the players comparison tests, during the test performing it was found that players have a similar performance and the home/away effect is not significant.

The average shot distance has shown to be linked to the empirical probability of the players for being successful making a shot.

References

Some consulting materials used for this work can be found below. See for example (Pishro-Nik 2014) and (Team 2020).

Pishro-Nik, Hossein. 2014. *Introduction to Mathematical Statistics*. First. Kappa Research LLC.

Team, R-Core. 2020. *R Language Definition*. Fourth. R-Core Team.