Spring 2016 Semester

STAT 3080: From Data to Knowledge - Final Project

# An Analysis of Steph Curry's Performance During 2014-2015 NBA Seasons

Zihan Ni (zn8ae), Zhiwei Zhang(zz3px)

Pledge:

# Introduction

**Data Overview:**

**Data:** Stephen Curry 2014-15 Game Log

**Source:** Basketball-reference.com

http://www.basketball-reference.com/players/c/curryst01/gamelog/2015/

**Who is Steph Curry?**

Wardell Stephen "Steph" Curry II (born March 14, 1988) is an American professional basketball player for the Golden State Warriors of the National Basketball Association (NBA). He is considered by some to be the greatest shooter in NBA history. Curry won the 2015 NBA Most Valuable Player (MVP) award and is a three-time NBA All-Star. (Wikipedia)

**Specific data related terminology:**

The game-log data was recorded by NBA staff during game time. It records all the game statistics of Steph Curry, a player of the Golden State Warriors, during NBA 2014-2015 season. The data was collected and released by Basketball-reference.com.

**Variables:**

(Boolean)Home: If a game is played at another team's home court, we use @ to represent. Empty means it is the Golden State Warriors' home game.

(String)Opp: Opponent Team Abbreviates. For example, Golden State Warriors -> GSW

(Integer)Winning: Winning points – the points Warriors get minus the points the opponent team gets at the end of the game.

(Integer)FG: Field Goal Made – total baskets made by Steph Curry, only including ones that go into the basket.

(Integer)FGA: Field Goal Attempts - the number of times that Curry attempted shooting the ball regardless of its getting into the basket or not.

(Quotient)FG%: Field Goal Percentage (FG/FGA) – the percentage of field goal attempts that results in scoring calculated by FG/FGA.

(Integer)3P: 3-Pointer Made – total number of 3-pointer shots made by Steph Curry, only including ones that go into the basket.

(Integer)3PA: 3-Pointer Attempts - total number of times that Curry attempted shooting from the 3-point range regardless of its getting into the basket or not.

(Quotient)3P%: 3-Pointer Percentage (3P/3PA) – the percentage of 3-pointer shot attempts that results in scoring calculated by 3P/3PA

(Integer)FT: Free Throws Made – total number of times that Curry made a free throw. In basketball, free throws or foul shots are unopposed attempts to score points from a restricted area on the court

(Integer)FTA: Free Throws Attempts – total number of times that Curry attempts to shoot a free throw. In basketball, free throws or foul shots are unopposed attempts to score points from a restricted area on the court

(Quotient)FT%: Free Throw Percentage (FT/FTA) – the percentage of free shot attempts that results in scoring calculated by FG/FGA

(Integer)TRB: Total Rebounds - total number of times that Curry retrieves the ball after a missed field goal or free throw.

(Integer)AST: Total Assists – total number of times that Curry passes the ball to a teammate in a way that leads to a score

(Integer)STL: Steals - In basketball, a steal occurs when a defensive player legally causes a turnover by his positive, aggressive action(s). (Wikipedia)

(Integer)TOV: Turnovers - In basketball, a turnover occurs when a team loses possession of the ball to the opposing team before a player takes a shot at his team's basket. （Wikipedia）

(Integer)PTS: Points – total points Curry gets in a specific game

(Integer)+/-: Efficiency – calculated according by NBA using particular defined formula to determine the contribution of players in the game. The formula is: (PTS + REB + AST + STL + BLK − Missed FG − Missed FT − TO)(NBA.com)

**Details of important variables:**

**Winning:**

Mean = 10.5625

Standard Deviation = 12.07555

**FG:**

Mean = 8.1625

Standard Deviation = 2.952981

**FGA:**

Mean = 16.7625

Standard Deviation = 4.237398

**FG%:**

Mean = 0.4834625

Standard Deviation = 0.1092803

**3P%:**

Mean = 0.4353375

Standard Deviation = 0.1845962

**FTA:**

Mean = 4.2125

Standard Deviation = 2.795538

**AST:**

Mean = 7.7375

Standard Deviation = 2.763889

**TOV:**

Mean = 3.1125

Standard Deviation = 1.807115

**PTS:**

Mean = 23.75

Standard Deviation = 8.155917

**+/-:**

Mean = 11.4875

Standard Deviation = 11.85994

**Basic graphs of each variable: Appendix(G1)**

**Steph Curry's Field Goal Percentage & three point percentage: Appendix(G2)**

**Steph Curry's Efficiency/Total points per match/Total Turnover per match comparison: Appendix(G3)**

**Steph Curry's Efficiency v.s Team Winning: Appendix(G4)**

## Results & Analysis

**Hypothesis Testing:**

**One-sample test: One-Sample t-test**

**Variable used:**

We are interested in if Curry's FGA is among the highest in NBA.

**Why:**

According to the statistics of Average Field Goal Attempts Per Game of NBA Season 2014-2015, the mean of the top 50 highest FGA per game data is about 15 (SportingCharts(FGA)). The rationale here is that the more a player's Field Goal Attempts is, the more the basketball team is relying on this player because the fact that the person execute the Field Goal Attempts is the deciding factor of if a team can score in this round of offence.

**Purpose of the test:**

We will test if the mean of Curry's FGA is greater than 15. The result of this test would tell us if the Golden State Warriors is relying on Curry to win their games.

**Assumption:**

Based on the histogram and qqplot of FGA,the distribution of FGA is symmetric and normal(Appendix(G2)), we can assume normality. The population standard deviation of Curry's FGA is unknown, so we are using one sample t-test  here instead of one sample z-test.

**Hypothesis:**

$$Ho: \mu = 15 \text{ versus } Ha: \mu > 15$$

H0: Curry's average Field Goal Attempt per game is 15.

Ha:  Curry's average Field Goal Attempt per game is greater than 15.

**Result:**

We have t statistic with degree of freedom 79 equals to 3.7203. We have p-value equals to 0.0001854, which is smaller than 5%, thus we can reject null hypothesis. We have significant evidence to assume that the true mean of Curry's FGA per game is greater than 15. In terms of our data,the one sample t-test indicates that the average number of times that Curry attempted shooting the ball regardless of its getting into the basket or not is greater than 15.

**Two-sample dependent test: Paired t-test**

**Variable used:**

The variables pair FG%, Field Goal Percentage and 3P%, 3-Pointer Percentage will be tested.

**Why:**

Since Steph Curry is famous for his 3-point shooting skill (CBSSports). We want to know if Curry is so good that the mean of his 3-point shooting percentage(3P%) surpasses his overall shooting percentage(FG%).

**Purpose of the test:**

If the data indicates that his 3P% is greater than his FG%, then it would be better if Curry attempts 3-point every time he wants to score. If not, we can say that it would be better if Curry tries from a diverse set of ranges.

**Assumption:**

Based on the histogram and qqplot of 3P% and FG%, the distributions are approximately symmetric and normal(Appendix(G2)), so we can assume normality. And they have the same number of data entries, and since Field Goal Percentage includes 3-Pointer Percentage, the two variables are dependent. The population standard deviations are also unknown here. Hence, we use the paired t-test.

**Hypothesis:**

$$Ho: \mu x = \mu y \text{ versus } Ha: \mu x > \mu y$$

Ho: The average of Curry's three-point percentage(x) is the same as his average field goal percentage(y).

Ha: Curry has higher average three-point percentage(x) than average field goal percentage(y).

**Result:**

We have paired t test statistic with degree of freedom 79 equals to -2.9056. We have p-value equals to 0.9976, which greater than 5% thus we can fail to reject null hypothesis. We don't enough evidence to say Curry has higher average three-point percentage than average field goal percentage. In terms of our data, we can't conclude that Curry has higher 3-point shooting percentage(3P%) than his overall shooting percentage(FG%), even though Curry is good at three-point shooting.

**Two-sample independent tests:**

**1. F-test for variances**

**Variable used:**

What will be tested next are two samples taken from the PTS attribute. Sample pts1 are the PTS data from match 1-40 and pts2's data are from match 41-80.

**Why:**

During a typical NBA season, the tension of games will gradually increase as the schedule approaches the end of season because all teams want to compete for the few spots in the

playoff in order to compete for the champion. As the tension of games grows up, players tend to play more physically which lead to better defense in the game. According to statistics from Teamrankings, we observed that for most of the teams the opponent points allowed in the last few games are lower than the average points allowed for the entire season.(Teamrankings) We want to know if Curry's scoring performance is affected by the increasing aggressiveness of defense or not.

**Purpose of the test:**

Since we want to know if Curry's scoring performance is affected by the increasing aggressiveness of defense or not. We can test if the variance of pts1 is smaller than the variance of pts2. If the result of our test indicates that the variance of two PTS data are equal, we can say that despite the fact that defense pressure has increased in the later half of season, Curry 's scoring performance is not affected. If not, we can say that since the variance of Curry's scoring increases in the late season, Curry's performance might be influenced by the increasing aggressiveness of defense.

**Assumption:**

Based on the histogram and qqplot of PTS, the distribution is approximately symmetric and normal(Appendix(G2)), so we can assume normality. All these matches are independent so our two groups of data are independent as well. Hence, we use the F-test for variances, testing if the variance of pts1 is smaller than the variance of pts2.

**Hypothesis:**

$$Ho: \sigma^2 x = \sigma^2 y \text{ versus } Ha: \sigma^2 x < \sigma^2 y$$

Ho: variance of pts1(x) data is the same as that of pts2(y)

*Ha: the variance of pts1 is smaller than the variance of pts2*

**Result:**

We have F statistic with degree of freedom 39 equals to 0.58895. P-value for this test is 0.05122, which is greater than 5%, thus we fail to reject our null hypothesis at 5% significance level. Hence, we don't have sufficient evidence to conclude that the variances of the two groups of Curry's total points are different. In terms of our data, at 5% significance level, we

can't conclude that the variance of Curry's scoring performance increases in the late season and can't not state that there is inconsistency in Curry's performance. However, since our p-value is only 0.001 above 5%, at a looser significance level such as 10%, we can still reject our null hypothesis, and say that Curry's scoring performance varies across the season. One may say that there is some degree of variance increase, but not a very significant amount.

**2. Wilcox Rank-Sum test**

**Variable used:**

What will be tested next are two samples taken from the FTA attribute. Sample FTA1 are the FTA data from match 1-40 and FTA2's data are from match 41-80.

**Why:**

During a typical NBA season, the tension of games will gradually increase as the schedule approaches the end of season because all teams want to compete for the few spots in the playoff in order to compete for the champion. As the tension of games grows up, players tend to play more physically which lead to better defense in the game. According to statistics from Teamrankings, we observed that for most of the teams the opponent points allowed in the last few games are lower than the average points allowed for the entire season.(Teamrankings)
We want to know if this increase of defending aggressiveness would lead to more Free Throw Attempts(FTA) for Curry since aggressive defense would create more fouls and more fouls may leads to more Free Throw by basketball rules.

**Purpose of the test:**

We want to know if the median of Curry's FTA increases due to the increasing aggressiveness of defense or not.
If the result of our test indicates that the median of two FTA data are equal, we can conclude that despite the fact that defense pressure has increased in the later half of season, Curry 's overall FTA is not affected.
If not, we can say that Curry's mean FTA increases due to the increase of late-season aggressive defense of opponent teams.

**Assumption:**

Based on the histogram and qqplot of FTA(Appendix(G2)), we can't assume normality. Hence, we use Wilcox Rank-Sum test for median. Additionally, two samples taken from the FTA attribute and each pair is chosen randomly and independently.

**Hypothesis:**

$$Ho: M_x = M_y \text{ versus } Ha: M_x < M_y$$

*Ho: The median of the FTA from matches 1-40(Mx) is the same as the median of FTA from matches 41-80(My)*

*Ha:* Curry's median of FTA from matches 41-80 is greater than median from matches 1-40

**Result:**

We have w statistic equal to 775.5. P-value for this test is 0.4063, which is greater than 5%, thus we fail to reject our null hypothesis. Hence, we don't have sufficient evidence to conclude that the median of FTA2 is greater than that of FTA1. In terms of our data, there is not enough evidence from the Wilcox Rank-Sum test indicating Curry's FTA increasing in the later half of season due to the increase of late-season aggressive defense of opponent teams.

**Categorical tests:**

**1. 1-sample proportion z-test**

**Variable used:**

Here we will test the proportion FG/FGA, field goal made over field goal attempt, which is actually the FG% variable.

**Why:**

A player with high FG/FGA value scores more if we fix the number of Field Goal Attempt. Hence, having an higher FG/FGA means a player is efficient.

According to the statistic at SportingCharts.com, to rank as top 200 players with highest Field Goal Percentage one has to have a FG/FGA equal to or larger than 0.45. We take top 200 here

to represent a high level of FG/FGA because in basketball there are different positions and some positions such as Center and Forward have comparatively higher FG/FGA because they usually attack basket at a closer distance than Guards such as Curry due to their bigger and taller body size. Hence, to be fair we take all positions into consideration and take the top 200 highest FGA players. (SportingCharts(FG%))

We believe that as a NBA star and an efficient player, Curry should have a FG/FGA higher than 45%. Thus, We want to know if Curry's FG/FGA value is greater than 45%.

**Purpose of the test:**

The purpose of this test is to see if Curry has a FG/FGA larger than 45%.
If the test result indicates that Curry has an FG/FGA > 45%. Then we can conclude that Curry is among the players with highest FG/FGA and hence he is an very efficient player.
If not, we do not have enough evidence to support that Curry is an efficient player so perhaps there are other ways to better justify the popularity of him and the high praises to him.

**Assumption:**

Based on the the histogram and qqplot of FG and FGA, the distributions are approximately symmetric and normal(Appendix(G2)). Also, for proportion test, we have n=80 and p=0.45, n*p=36,n*(1-p)=44, then we can assume normality.

**Hypothesis:**

$$Ho: p = 0.45 \quad versus \quad Ha: p > 0.45$$

*Ho: Curry has* a FG/FGA equal to 45%

*Ha*: Curry has a FG/FGA larger than 45%

**Result:**

We have test statistic x-squared equal to 7.3975. P-value for this test is 0.003266 which is smaller than 5% significance level. Thus we can reject our null hypothesis and say that Curry has a FG/FGA larger than 45%. In terms of data, we may say that Curry's proportion FG/FGA, which is the field goal made over field goal attempt is greater than 45% and conclude that Curry is among the players with highest FG/FGA and hence he is an very efficient player.

**2. 2-sample proportion test**

**Variable used:**

We want to test if the proportion of games won by Golden State Warriors and the proportion of the games in which Curry has positive efficiency are equal (both are with regards to all 80 games played in our data). Our variables are:

P1:*sum (curry$Winning > 0) / 80*

Proportion of games won by Golden State Warriors

P2: *sum(curry$`+/-` > 0) / 80*

Proportion of the games in which Curry has positive efficiency

**Why:**

We want to investigate if Curry's performance is a decisive factor of the number of games won by Golden State Warriors. Hence, we need to measure both the proportion of game in which Curry has positive efficiency and the proportion of game that Golden State Warriors won.

**Purpose of the test:**

By conducting the 2-sample proportion test we can gain knowledge about whether the proportion of game in which Curry has positive efficiency decides the proportion of game that Golden State Warriors won. If so, then it means that Curry's performance can be directly related to Golden State Warriors' performance. We might be able to use Curry's performance to predict Warriors' performance in this case. If not, then there should be other factors that contribute to determine the proportion of games won by Warriors, aka the performance of Warriors.

**Assumption:**

Based on the the histogram and qqplot of winning and efficiency `+/-`, the distributions are approximately symmetric and normal(Appendix(G2)). Also, for proportion test, we have n=80, p1=0.8375 and p2=0.7875. Therefore, n*p1=67,n*(1-p1)=13,n*p2=63,n*(1-p2)=17, thus we can assume normality.

**Hypothesis:**

$$Ho: p1 = p2 \quad versus \quad Ha: p1 \neq p2$$

*Ho:* the proportion of games won by Golden State Warriors and the proportion of the games in which Curry has positive efficiency are equal

Ha: the proportion of games won by Golden State Warriors and the proportion of the games in which Curry has positive efficiency are different

**Result:**

We have test statistic x-squared equal to 0.65641. P-value for this test is 0.4178, which is greater than 5%. Hence, we fail to reject null hypothesis and we don't have enough evidence to accept alternative hypothesis. In terms of our data, we can't conclude that the proportion of games won by Golden State Warriors and the proportion of the games in which Curry has positive efficiency are different. Then it means we might be able to relate Golden State Warriors' performance to Curry's performance in this case.

## Modeling

**Multiple Linear Regression(MLR):**

Training data for MLR - we randomly select 64 out of 80 tuples in our data to serve as our training data. We will later use the training data to generate predict interval

Testing data for MLR - we use the 16 out of 80 data other than training data to serve as our testing data. We will later use testing data to analyze our predict interval's accuracy.

**Description of model (notation) & variable used**

**Model:** (handwritten)

**Response Variable:**

(Integer)+/-: Efficiency – calculated according by NBA using particular defined formula to determine the contribution of players in the game. We already know that better performance in less time means higher efficiency, but we do not yet know what factors indicate better performance. Hence, we want to investigate what variables from our data contribute to determine the efficiency value of Curry.

**Explanatory Variables:**

**(Quotient)FG%: Field Goal Percentage (FG/FGA)** – calculated by FG/FGA, which is the number of shots curry made to the number of shooting attempts curry made. The higher shooting percentage means that Curry is making more goals if given a fixed number of shooting attempts. Since we want to investigate Curry's playing efficiency (or performance), we believe this variable would contribute positively.

**(Integer)PTS: Points** – total points Curry gets in a specific game. We do not think the total points Curry gets in a specific game is an effective criterion to measure performance. For example, it would be possible for Curry to make a large amount of shooting attempts in a single game but only make some of them. At last Curry's team loss because those attempts that Curry did not make are wasted and it would work better if Curry passed the ball to his teammates. If this is the case then Curry would have a very high total points but a bad performance. Hence, we want to see what happen if we remove this variable from our model.

**(Integer)AST: Total Assists** – total number of times that Curry passes the ball to a teammate in a way that leads to a score. Even though sometimes Curry does not make goals himself, he might pass the ball to a teammate who has better opportunity to score. Hence, we believe the fact that Curry make few attempt to score does not mean Curry does not have a good performance. If Curry passes a lot and his teammates score a lot, we can say that Curry as the facilitator of successful scoring still perform well. Hence, we believe this variable contribute positively to efficiency.

**(Integer)TOV: Turnovers** - In basketball, a turnover occurs when a team loses possession of the ball to the opposing team before a player takes a shot at his team's basket. (Wikipedia) Since losing possession of the basketball means giving the opponent a chance to score. We believe that more turnovers means worse performance and hence contribute negatively to the efficiency value of a player.

**Intersections:** we include all intersections of these explanatory variables in our maximum model because we think that since there might be 3 or 4 variables are significant hence it might be possible that some intersections are significant as well. However, the maximal model might not be the best model so we will test its significance and try to reduce it to simpler and stronger form.

**Coefficients Variables $\beta$ i:**

$\beta_0$ is our intercept. In graph if all the explanatory variables are 0 then we would have our efficiency value +/- = $\beta_0$ It is not hard to guess that $\beta_0 = 0$ since if a player does not have any stat in a game then only the trivial value 0 can represent his performance - neither good nor bad.

Other $\beta$ i's are slopes. They would tell us what would happen to our response variable +/- when one unit of increase happen to the explanatory variable with this coefficient.

**What would this model tell us:**

Our model equation would tell us the magnitude each of the explanatory variables, namely FG%, PTS, AST, TOV and their intersections, affect the efficiency of Curry represented by variable +/- and whether each one of them positively affects +/- or negatively affects +/- or does not affect +/- at all.

**Summary of model-simplifying process**

1. We would first need to produce the ANOVA table to see if the explanatory variables

collectively have a statistically significant effect on +/-. In other words, we need to test if our model is significant overall.

Our assumption is:(handwritten)

They would tell us if the coefficients of our explanatory variables are 0 or not. In other words, does any of our explanatory variables affect our response variable (+/-)?

If we reject H0, the model is significant, which means at least one of the explanatory variable has a non-zero coefficient. Hence, we know that our model explains +/- to some degree.

If we fail to reject H0, then our model is not significant. If our model is not significant then we would not go into the simplify process and instead we just need to find another model.

2. If we rejects H0 in the above test, we would go on and conduct the T-test for all our slopes(coefficients). If we find a slope of a variable to not be significant then we have no need to include it in our model.

Our assumption is:(handwritten)

They would tell us if some of the coefficients are 0. Having a regression coefficient of 0 means any change in the explanatory variable results in no change in the response variable. So that we can remove those variable with 0 as coefficient and hence simplify our model.

If we reject H0 for some $\beta$ i, it means that the coefficient $\beta$ i is not 0 and hence the variable with this coefficient is significant.

If we fail to reject H0 for some $\beta$ i, it means that the coefficient $\beta$ i is 0 and hence the variable with this coefficient can be removed without affecting the overall significance of our model too much.

3. We remove the variable with the highest p-value that does not reject H0 if we have multiple p-value > $\alpha$, our significance level. Order of removal: In order to get the simplest result as possible, we consider the most complex interaction(one which contains more number of variables)first. Hence, we remove the element with highest p-value at the most complex level.

4. We repeat the entire process until all of our explanatory variables are very significant and hence have a p-value > $\alpha$.

**Variables removed in order:**

train$`FG%`:train$PTS:train$AST:train$TOV

train$`FG%`:train$PTS:train$AST

train$`FG%`:train$PTS:train$TOV

train$PTS:train$AST:train$TOV

train$`FG%`:train$AST:train$TOV

train$AST:train$TOV

train$PTS:train$TOV

train$PTS:train$AST

train$`FG%`:train$AST

train$`FG%`:train$PTS

train$`FG%`:train$TOV

train$PTS

**Assumptions need to be checked:**

1. All of our variables are quantitative. Our outcome variable is quantitative and continuous.

2. Check residual assumptions:

**Homoscedasticity:** residuals versus order of observation plot shows constant residual variance. (Appendix:G6)

**Normal distribution of error:** use quantile-quantile plot to test normally distributed error, the linear pattern indicates no significant departure from normality. (Appendix:G5)

**Independence of error:** the residuals versus order of observation plot shows that the distribution of residuals has no association with the order of the observation.(Appendix:G6) Moreover, the time series plot shows that the distribution is not affected by time. (Appendix:G7)

**Final Sample Equation:**

+/- = -6.5236 + 25.0483(FG%) + 1.3010(AST) - 1.6528(TOV)

**Interpretation of the model:**

If all FG%, AST and TOV are 0, +/- would be -6.5236. Originally the efficiency would be 0 if all performance statistics are 0. However, we are only testing FG%, AST and TOV so there might be other factors that contribute to determine +/-.

One unit of increase in FG%, while holding other factors fixed, will result in 25.0483 increase in +/-

One unit of increase in AST, while holding other factors fixed, will result in 1.3010 increase in +/-

One unit of increase in TOV, while holding other factors fixed, will result in 1.6528 decrease in +/-

**Detail and analysis of how we simplified from the second to last model to the final model using the training data:**

Our second to last model includes variables FG%, PTS, AST and TOV using the training data.

We conduct linear regression on these four variables and use t test on the slopes of these individual variables. If we find the slope of a variable to be not significant then it means that we have no need of it in the model. We test null hypothesis if the slope of a variable is zero. The hypothesis is: *Ho*: $\beta$ j = 0. From the model output, we have t-statistic equal to -0.937 and p-value equal to 0.3582 for variable PTS. Therefore, we fail to reject our null hypothesis at 5% significance level. In MLR this means that the slope of variable PTS is 0, meaning we have no need of it in the model.

A new model is developed by removing PTS variable. The MLR output from the new developed model have all three variables FG%, AST and TOV significant with p-value smaller than 5%. Also, based on the 5.178 F-statistic and 0.003005 p-value, we can conclude that these three variables collectively have a significant effect on Curry's efficiency

After conducting linear regression on these four variables, we found that only variable PTS was not significant with p-value equal to 0.3582, so we can reject the null hypothesis at 5% significance level and remove variable PTS from our model. Hence, our final model includes variables FG%, ATS and TOV.

**Subjective conclusion of model:**

**Meaning of final model:**

In our final model, the three factors that determines +/- are FG%, AST and TOV. This means FG%, AST and TOV are factors that affect Curry's efficiency.

FG% contribute positively to +/-. This means that the higher Field Goal Percentage Curry has, the more efficient he is. This is reasonable since an higher shooting percentage means that Curry is making more goals if given a fixed number of shooting attempts.

AST contribute positively to +/- as well. This means that the more assists Curry has, the more efficient he is. This is reasonable because assist is count when Curry's passing to ball to others and helping other teammates score.

TOV contribute negatively to +/-. This means that the more turnover Curry has, the less efficient he is. This is reasonable because if Curry has a turnover it means that Curry handed the scoring opportunity to the opponent team while contribute nothing positive to his own team.

Hence, in our final model, the higher FG%, the more AST, and the less TOV Curry has, the more efficient he is.

**How well our model predicts values in testing dataset:**

After we conduct predict on our final model and plot the actual efficiency from the testing dataset, we find that all these 16 testing data fall within the prediction interval (Appendix G9). Hence, we may say that our final forecasting model is valid and can effectively predict Curry's efficiency.

**Other statistics that helps our analysis**

**Sum of Squared Residuals (SSE)**

This would tell us the degree of scatter of our model. The SSE for our final model is 6624.092, which means the sum of the squared differences between the prediction for each observation and the population mean is 6624.092.

**Coefficient of Determination (R^2)**

This would tell us the fraction of the total variation in our response variable that is explained by variation in explanatory variables.

The Coefficient of Determination for our final model is about 0.2057. Hence we may conclude that about 20.57% of the total variation in our response variable that is explained by variation in explanatory variable. The reason for this is that FG%, AST and TOV are not all the factors that determines Curry's effectiveness(+/-). According to the NBA official definition, the +/- variable is not merely determined by FG%, AST and TOV. However, these three variables do play important roles in determining +/-.

**Time Series Analysis**

**Description of variable:**

(Integer)+/-: Efficiency – calculated according by NBA using particular defined formula to determine the contribution of players in the game. We want to use time series analysis to obtain statistical measures of Curry's efficiency and, if possible, we want to figure out the trend or periodicity of Curry's efficiency.

**Training data for Time Series Analysis** - we select the first 72 out of 80 (first 90%) tuples in our data to serve as our training data. We will later use the training data to predict the last 8 tuples and analyze the accuracy of our prediction model.

**Testing data for Time Series Analysis** - we use the last 8 (last 10%) data to serve as our testing data. We will use the prediction model we build to predict the range of these last 8 tuples and then analyze our predicted interval's accuracy.

**Time Series data plot and patterns:**

Time Series plot: Appendix(G10)

Patterns: From observation we notice that there seems to be no significant general trend in this times series graph. As for seasonality, we are unclear if the graph contains seasonality to some degree. We need to check ACF and PACF plot to confirm our thoughts about trend and seasonality.

**Autocorrelation and Partial Autocorrelation plots:**

ACF plot: Appendix(G11)

The ACF plot of the time series gives correlation between x at time t and x at time t-h, for h = 1,2,3 .etc. Theoretically, autocorrelation between x at time t and at time t-h is calculated as : covariance(Xt,X (t-h))/ variance(Xt) . Autocorrelation tells us how a population at one time point t is related to a previous time point (t-h) h points part.

From ACF plot of Curry's efficiency, it is clear that there is only one significant spike at time 0 and after lag 0 it quickly decays to 0. We may say that the data should be stationary with regard to this evidence. However, we can still transform the data using detrending or differencing techniques later.

PACF plot: Appendix(G12)
Partial autocorrelation describes the relationship between a time point t and the population at lag h once we have controlled for the correlations between all of the successive time points between t and time point t-h.

From PACF plot of Curry's efficiency shows 0 significant spike and no significant trend. Hence, we cannot make any conclusion at this point with regard to the PACF graph. We may need further processing before we start to build our model.

**Stationarity:**
**Why helpful:** We generally prefer to analyze a stationary sequence in time series analysis because a *stationary* time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Those characteristics of a stationary time series allows us to better estimate autocorrelation and other quantities based on the assumption that the time series can be rendered approximately stationary. In terms of our data, we could predict Steph Curry's efficiency in future games once we have a stationary efficiency time series.

**How to check:**
We use ACF to check stationarity. An ACF which drops to 0 very quickly is indicative of a stationary series.

**When we do not have stationarity**:

We will need to reduce the effects of nonstationarity. The most common two ways to make a time series stationary are detrending with linear regression and differencing. To conduct detrending technique, we will first obtain an estimate of the trend component and work with the residuals; then subtract the trend component from the observed values which gives residuals. However, sometimes, detrending is not enough to transform a time series, another method we can use to obtain a stationary process is to difference the data. The first difference is the series of change from one period to the next. After differencing, it is helpful to look at the time plot, as well as the ACF for identifying stationary series. In terms of our data, we use the second technique to transform the data because based on the time series plot of Curry's efficiency, there is no obvious decreasing or increasing trend component.

**Smoothing for long term trend analysis:**

If we want to know Curry's long-term performance trend, we need to use some smoothing technique to smooth our data so that we can better analyze the trend. Smoothing can help us dampen irregularities so we got a clearer idea on the time series. Smoothing is used to smooth out the short term random fluctuations so that long term trends can be emphasized. Here we will use kernel smoothing to smooth our data.

**Kernel Smoothing:**

Kernel smoothing contributes to the estimate of the smooth function at a point t from local points decline. Two choices for K are standard normal and uniform. Here we will use standard normal kernel smoothing for our data and we focus on choosing a right bandwidth because choice of kernel is not as important as choosing a bandwidth.

The idea of the kernel average smoother is the following. For each data point $X_0$, choose a constant distance size $\lambda$ (kernel radius, or window width for p = 1 dimension), and compute a weighted average for all data points that are closer than $\lambda$ to $X_0$ (the closer to $X_0$ points get higher weights).

We will try bandwidth 3, 6, 12 for our data. The corresponding graphs are Appendix(G13), Appendix(G14) and Appendix(G15)

It is not hard to observe that when we use bandwidth 3, the smoothing is not very effective and there are still a lot of noise. Hence, if we use bandwidth 3 we might under smooth. As for bandwidth 12, the smoothing effective is too much so that the desirable variation is obscured. If we use bandwidth 12, we might over smooth our data. If we look at the kernel smoothing by bandwidth 6, we can see the line reflects the general trend while smooth out some undesirable noise. Hence, we will adapt a kernel smoothing with bandwidth 6.

We can see from our smoothing line that Curry's efficiency fluctuates before game number 40. During the games indexed from 40 to 42, there seems to be a big drop of his performance. After game 42, his overall performance level gradually and stably grew. We can see that at the last game he had an 0 efficiency. This was due to the fact that he did not play at all in that game.

**Seasonality analysis:**

From the ACF (Appendix:G11) and PACF(Appendix:G12) graphs of Curry's efficiency data, there seemed to be no seasonality in Curry's performance because the ACF drops to 0 after lag0 and did not have any significant value again while the PACF does not have any significant spike. Moreover, it does not seem like the ACF and PACF plot contains any recurring trend or shape of data. So we do not need to use spectral analysis to deal with seasonality here.

**Model Building:**

If we use our original data we will have a ACF plot indicating q = 0 and PACF plot indicating p=0. In this case it means our data is random walk and it would be hard to build a model to make prediction. Hence, we try to take the difference of the data to see if we can find a model to fit the data after differencing.

**EDA Analysis:**

We check for stationarity in a time series after differencing(Appendix G15) if there is still any trend, increasing variability and seasonality. In terms of our efficiency plot, it suggests the mean and variance are relatively constant. Then we use ACF and PACF plots together to identify models. We use ACF plots to get q,Q(early lags for q, multiples of s for Q) for AR(p) Models; we use PACF plots to get p,P(early lags for p, multiples of s for P) for MA(q) Models; ARMA(p,q) models have ACF and PACF that both decay exponentially to 0. Based on our ACF plot of transformed efficiency (Appendix G16), ACF drops to 0 after lag 1, so our best guess of q is 1. Based on our PACF plot of transformed data(Appendix G17), our best guess of p is 3.

Based on ACF and PACF plots, we would have an idea about the values of p and q. Hence my suggestions on possible AR, MA and ARMA models are: MA(1), AR(3) and ARMA(1,3). Then, we want to investigate the model estimation and diagnostics for these three models.

**Model Diagnostics:**

**MA(1) Model:**

Time series plot of residuals:Appendix(G18)

We check if the variance is constant using a time series plot of the residuals. The time series plot of residual plot shows no significant trend or seasonality, which indicates that the variance is constant. We double check by plotting the ACF of residuals next.

ACF of residuals:Appendix(G18)

According to our graph, there is no spike after lag 0, which means the variance is relatively constant.

Ljung-Box-Pierce statistic:

We conduct Ljung-Box-Pierce test to see If the p value is greater than 0.05 then the residuals are independent which we want for the model to be correct. The p-value for the MA(1) model is 0.3629, which is desirable.

**AR(3) Model:**

Time series plot of residuals:Appendix(G19)

We check if the variance is constant using a time series plot of the residuals. The time series plot of residual plot shows no significant trend or seasonality, which indicates that the variance is constant. We double check by plotting the ACF of residuals next.

ACF of residuals:Appendix(G19)

According to our graph, there is no spike after lag0, which means the variance is relatively constant.

Ljung-Box-Pierce statistic:

We conduct Ljung-Box-Pierce test to see If the p value is greater than 0.05 then the residuals are independent which we want for the model to be correct. The p-value for the AR(3) model is 0.6839, which is desirable.

**ARMA(3,0,1) Model:**

Time series plot of residuals:Appendix(G20)

We check if the variance is constant using a time series plot of the residuals. The time series plot of residual plot shows no significant trend or seasonality, which indicates that the variance is constant. We double check by plotting the ACF of residuals next.

ACF of residuals:Appendix(G20)

According to our graph, there is no spike after lag0, which means the variance is relatively constant.

Ljung-Box-Pierce statistic:

We conduct Ljung-Box-Pierce test to see If the p value is greater than 0.05 then the residuals are independent which we want for the model to be correct. The p-value for the ARMA(3,0,1) model is 0.9209, which is desirable.

**Model Selection:**

Model diagnostics above suggest that all these three models may work. Hence, we compare their AIC and BIC in order to come up with the best model. Based on the output(Appendix G21), MA(1) has the smallest AIC and BIC value. Moreover, MA is simpler than ARMA model. Finally, the standard error of forecast of MA(1) model is the smallest. Therefore, our best model for variable efficiency would be MA(1).

**Future Forecast:**

To use our MA(1) model to make future forecast, we use Sarima.for() in R to carry out the forecast and use p,d,q values to forecast into the future. Since we are using MA(1) model, our p=0, d=0, q=1. In terms of our data, we use the first 72 values to train and we test the last 8 values.

The result of prediction (Appendix:G22) shows that our prediction interval is fairly large while our mean of prediction is just a straight line. This is due to the fact we are using MA(1) model and we are taking the first difference of our data. The resulting model is nearly random walk and hence it is hard to predict a largely random data. As we can see, if we plot the testing(last 10%) data in the graph, neither the mean nor the interval perfectly describe the testing data.

# Works Cited

SportingCharts:

(FGA):

Field Goal Attempts Per Game: 2014-15 NBA Season

http://www.sportingcharts.com/nba/stats/player-field-goal-attempts-per-game/2014/

(FG%):

Field Goal Percentage Leaders: 2014-15 NBA Season

http://www.sportingcharts.com/nba/stats/player-field-goal-percentage-leaders/2014/

CBSSports:

Harper, Zach. "Steph Curry 3-point Tracker: He's Still Hovering Just above 400 Makes."

http://www.cbssports.com/nba/eye-on-basketball/25543839/steph-curry-3-point-tracker-hes-still-hovering-just-above-400-makes

Teamrankings:

NBA Team Opponent Points per Game

https://www.teamrankings.com/nba/stat/opponent-points-per-game

# Appendix

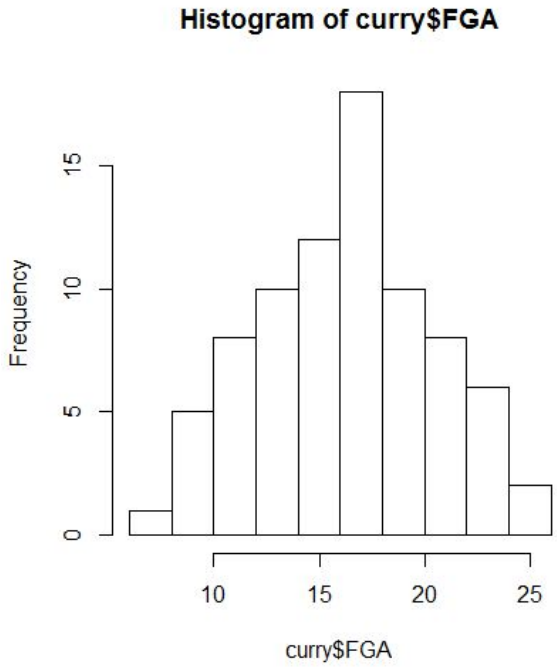
## Graphs:

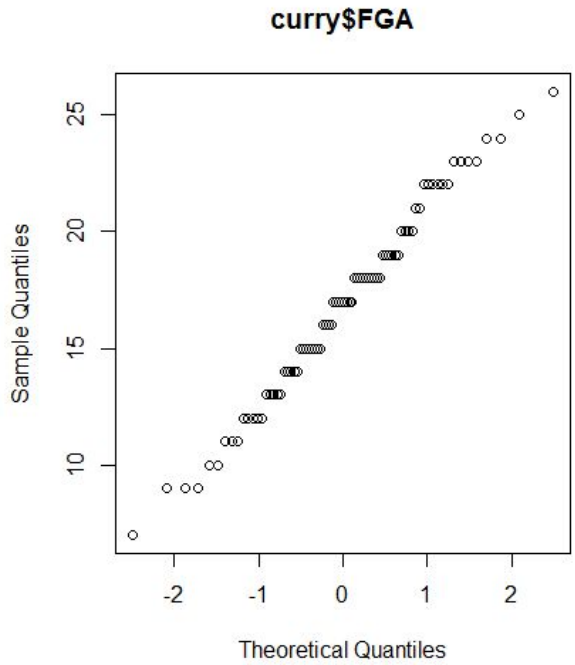
G0: first 10 observations:

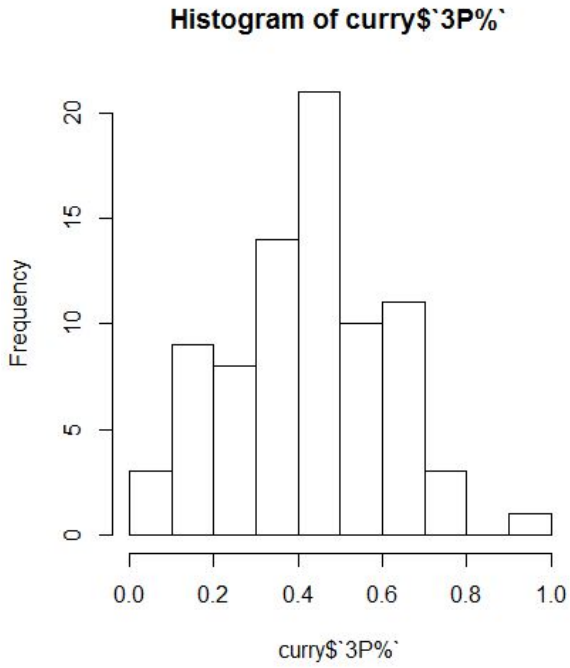
First 10 observations:

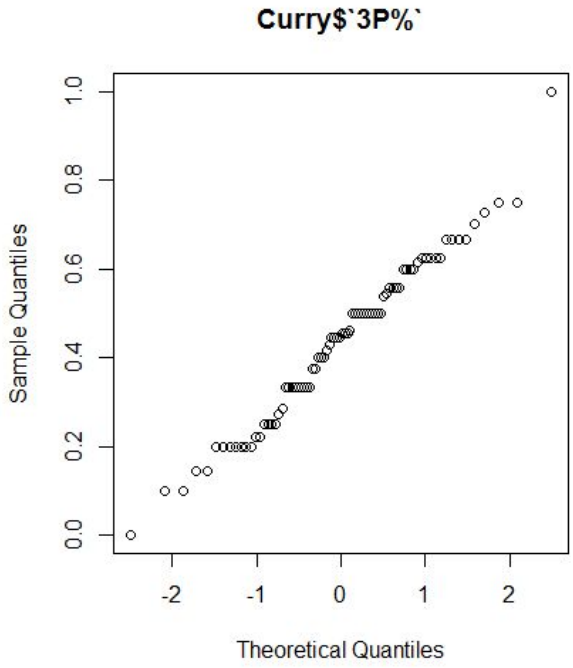
| | Home | Opp | Winning | FG | FGA | FG% | 3P | 3PA | 3P% | FT | FTA | FT% | TRB | AST | STL | TOV | PTS | +/- |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | @ | SAC | 18 | 7 | 17 | 0.412 | 2 | 9 | 0.222 | 8 | 9 | 0.889 | 10 | 5 | 6 | 4 | 24 | 20 |
| 2 | | LAL | 23 | 10 | 19 | 0.526 | 3 | 8 | 0.375 | 8 | 8 | 1.000 | 5 | 10 | 3 | 2 | 31 | 30 |
| 3 | @ | POR | 5 | 6 | 18 | 0.333 | 1 | 5 | 0.200 | 8 | 8 | 1.000 | 5 | 6 | 2 | 3 | 21 | 12 |
| 4 | | LAC | 17 | 9 | 18 | 0.500 | 4 | 8 | 0.500 | 6 | 6 | 1.000 | 6 | 7 | 1 | 5 | 28 | 21 |
| 5 | @ | HOU | 11 | 13 | 19 | 0.684 | 6 | 9 | 0.667 | 2 | 2 | 1.000 | 9 | 5 | 4 | 5 | 34 | 13 |
| 6 | @ | PHO | -12 | 10 | 20 | 0.500 | 4 | 10 | 0.400 | 4 | 4 | 1.000 | 2 | 10 | 5 | 10 | 28 | -6 |
| 7 | | SAS | -13 | 7 | 18 | 0.389 | 0 | 7 | 0.000 | 2 | 2 | 1.000 | 6 | 5 | 0 | 3 | 16 | -18 |
| 8 | | BRK | 8 | 6 | 12 | 0.500 | 3 | 7 | 0.429 | 2 | 2 | 1.000 | 3 | 5 | 0 | 3 | 17 | 6 |
| 9 | | CHO | 25 | 8 | 15 | 0.533 | 3 | 6 | 0.500 | 0 | 1 | 0.000 | 5 | 9 | 1 | 1 | 19 | 23 |
| 10 | @ | LAL | 21 | 10 | 19 | 0.526 | 5 | 9 | 0.556 | 5 | 5 | 1.000 | 4 | 15 | 1 | 3 | 30 | 27 |

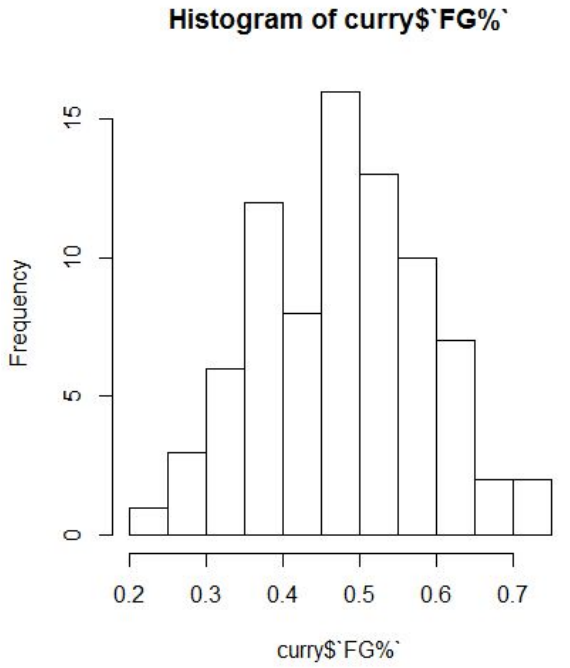
G1: Basic graphs of important variables

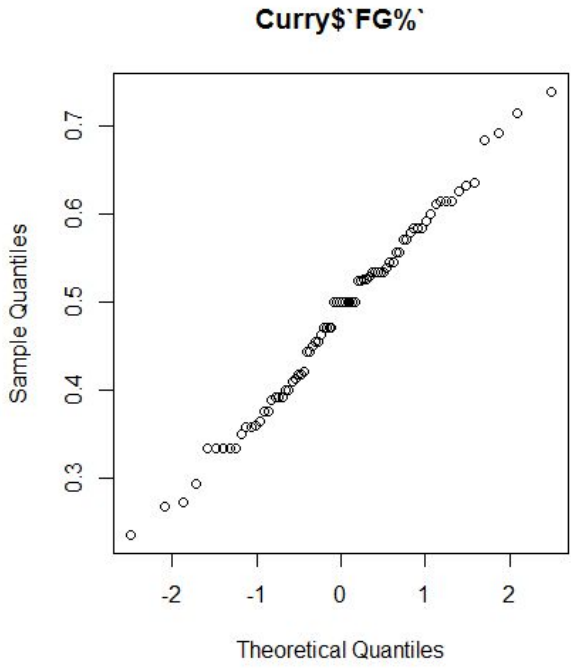


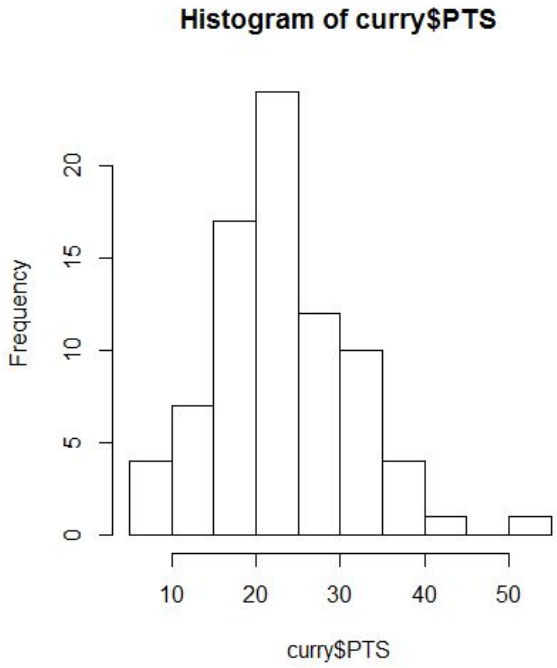



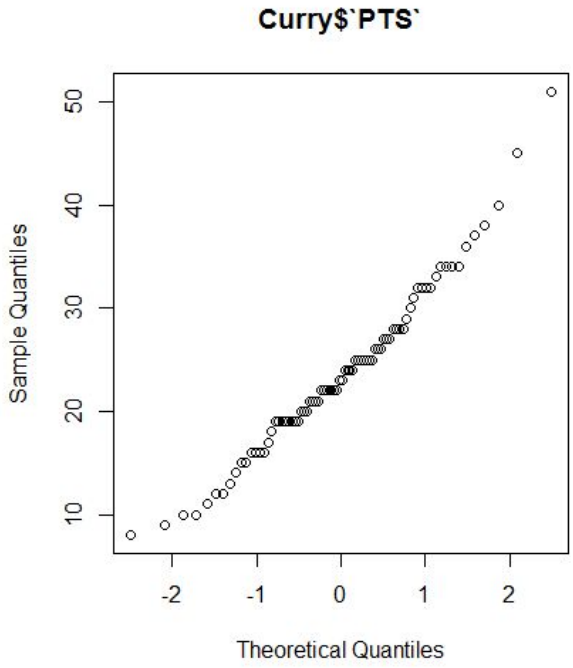

Histogram of curry$`3P%`
Frequency
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
curry$`3P%`

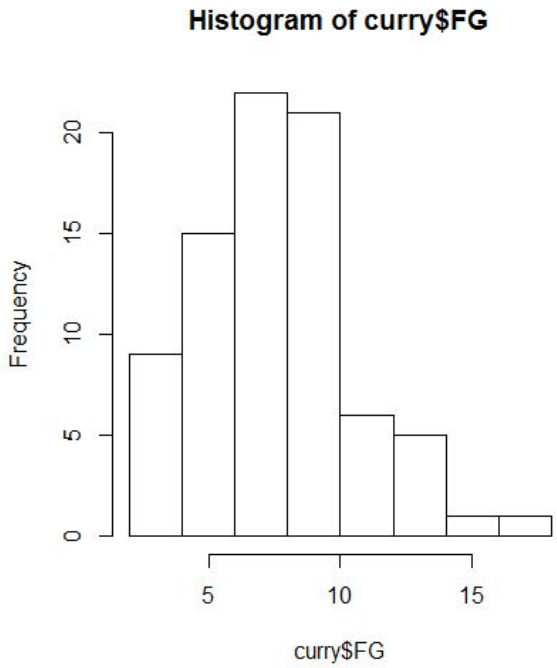

Curry$`3P%`
Sample Quantiles
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theoretical Quantiles

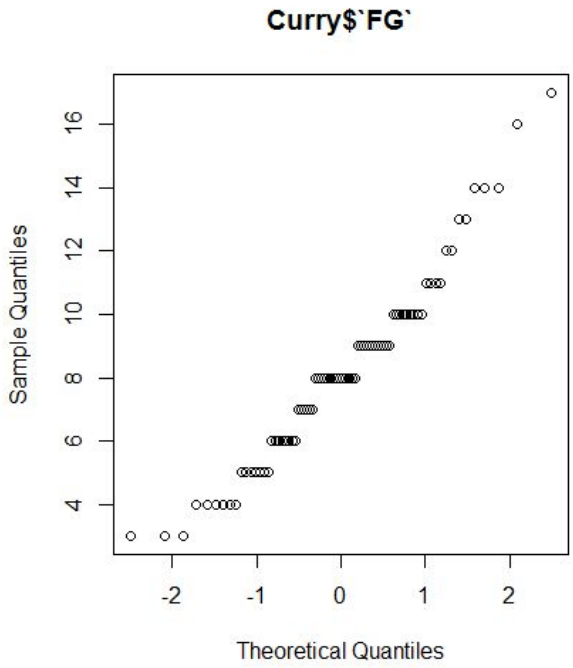

Histogram of curry$`FG%`
Frequency
0
5
10
15
0.2
0.3
0.4
0.5
0.6
0.7
curry$`FG%`

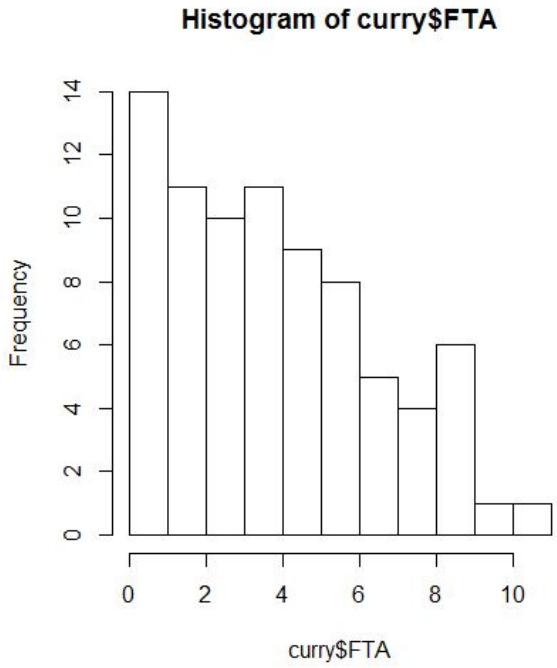

Curry$`FG%`
Sample Quantiles
0.3
0.4
0.5
0.6
0.7
-2
-1
0
1
2
Theoretical Quantiles

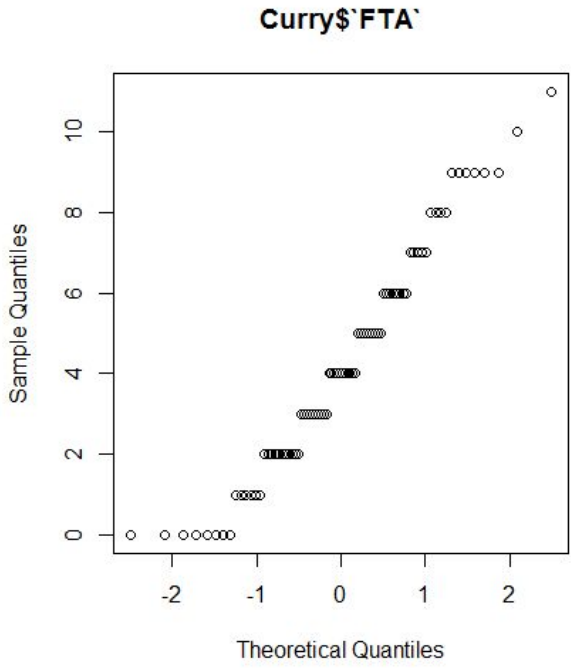

Histogram of curry$PTS
Frequency
0
5
10
15
20
10
20
30
40
50
curry$PTS

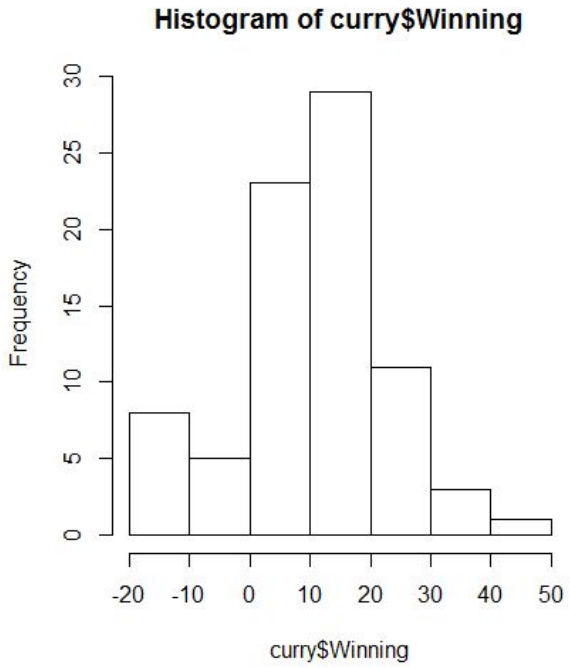

Curry$`PTS`
Sample Quantiles
10
20
30
40
50
-2
-1
0
1
2
Theoretical Quantiles

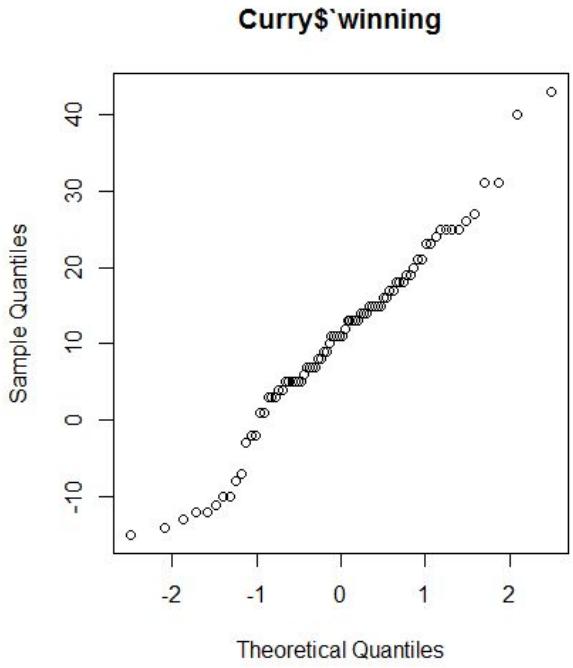

Histogram of curry$FG
Frequency
0
5
10
15
20
5
10
15
curry$FG

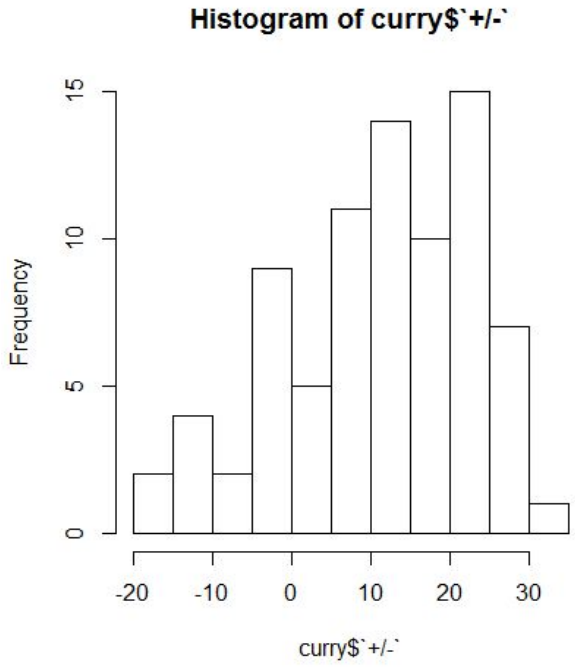

Curry$`FG`
Sample Quantiles
4
6
8
10
12
14
16
-2
-1
0
1
2
Theoretical Quantiles

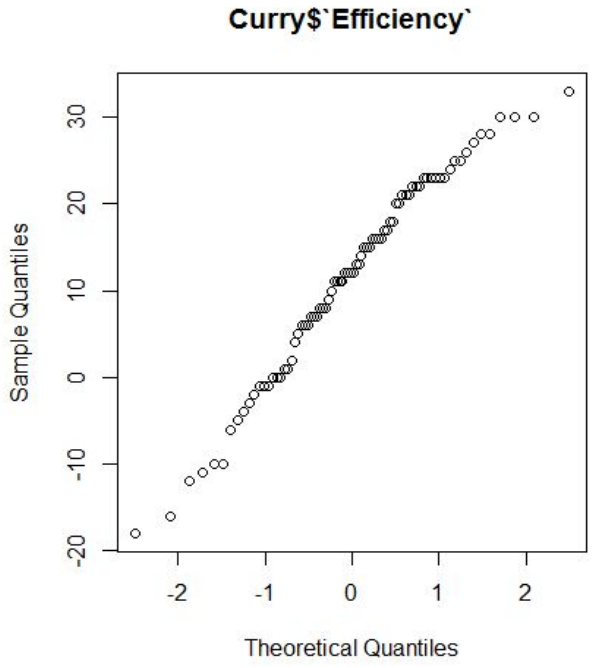

Histogram of curry$FTA
Frequency
curry$FTA
0
2
4
6
8
10
12
14

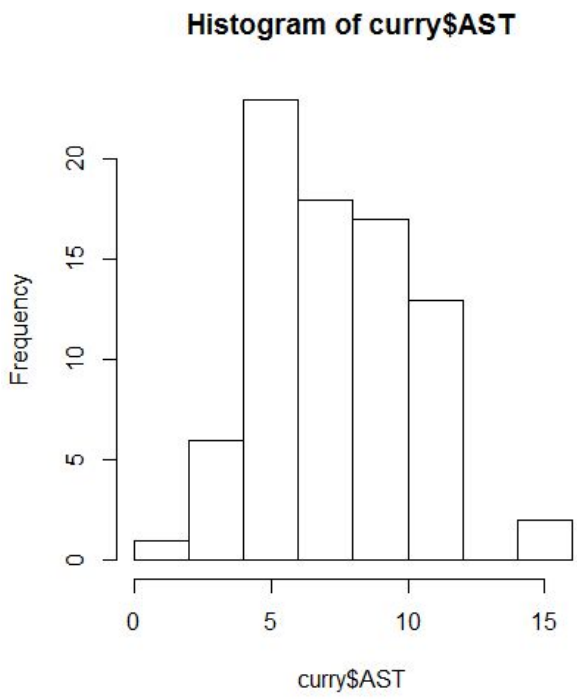

Curry$`FTA`
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2

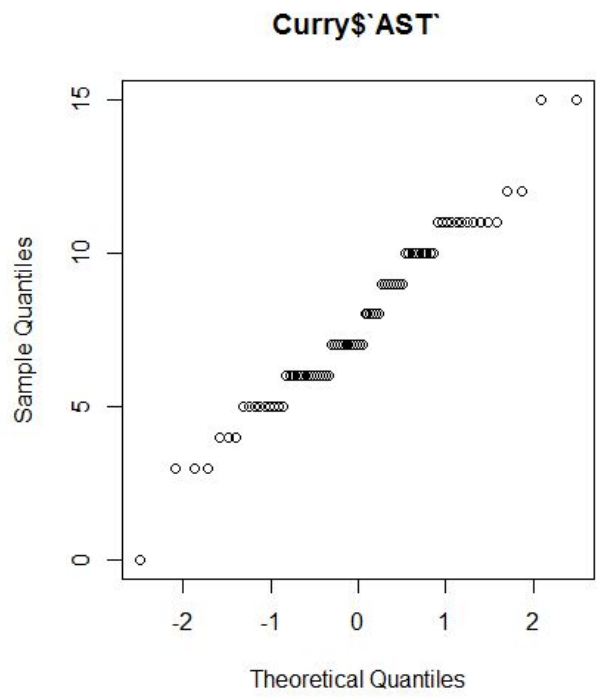

Histogram of curry$Winning
Frequency
curry$Winning
-20
-10
0
10
20
30
40
50

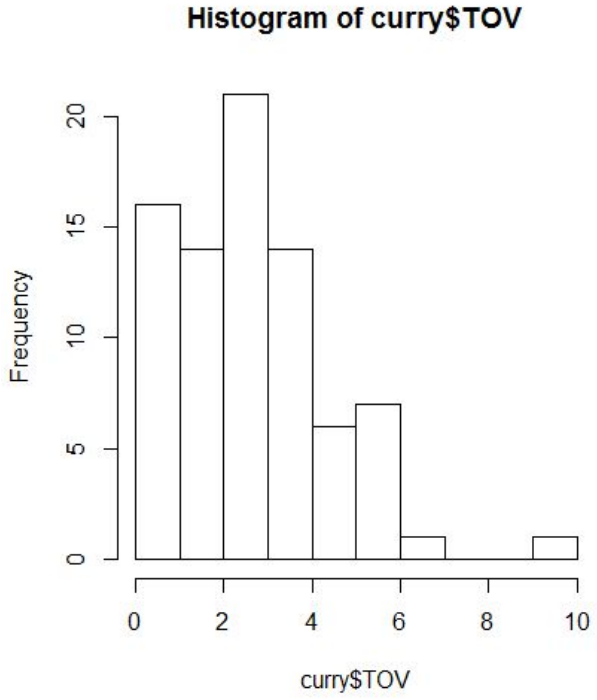

Curry$`winning
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2

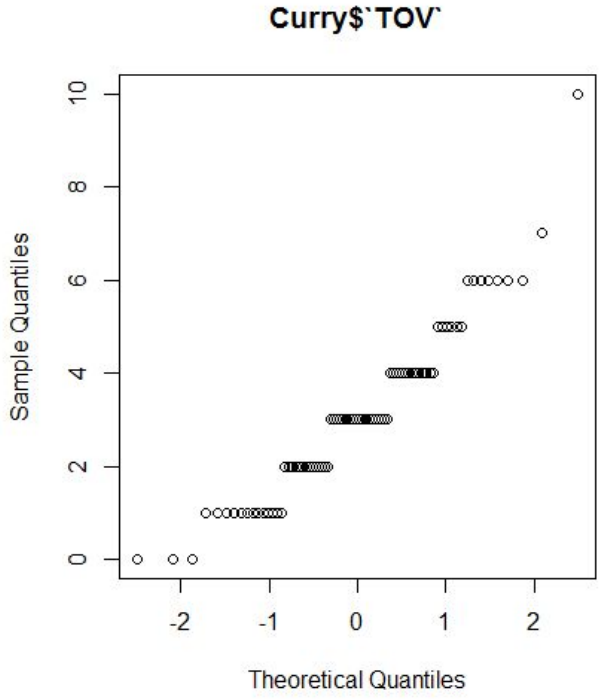

Histogram of curry$`+/-`
Frequency
0
5
10
15
-20
-10
0
10
20
30
curry$`+/-`


Curry$`Efficiency`
Sample Quantiles
-20
-10
0
10
20
30
-2
-1
0
1
2
Theoretical Quantiles

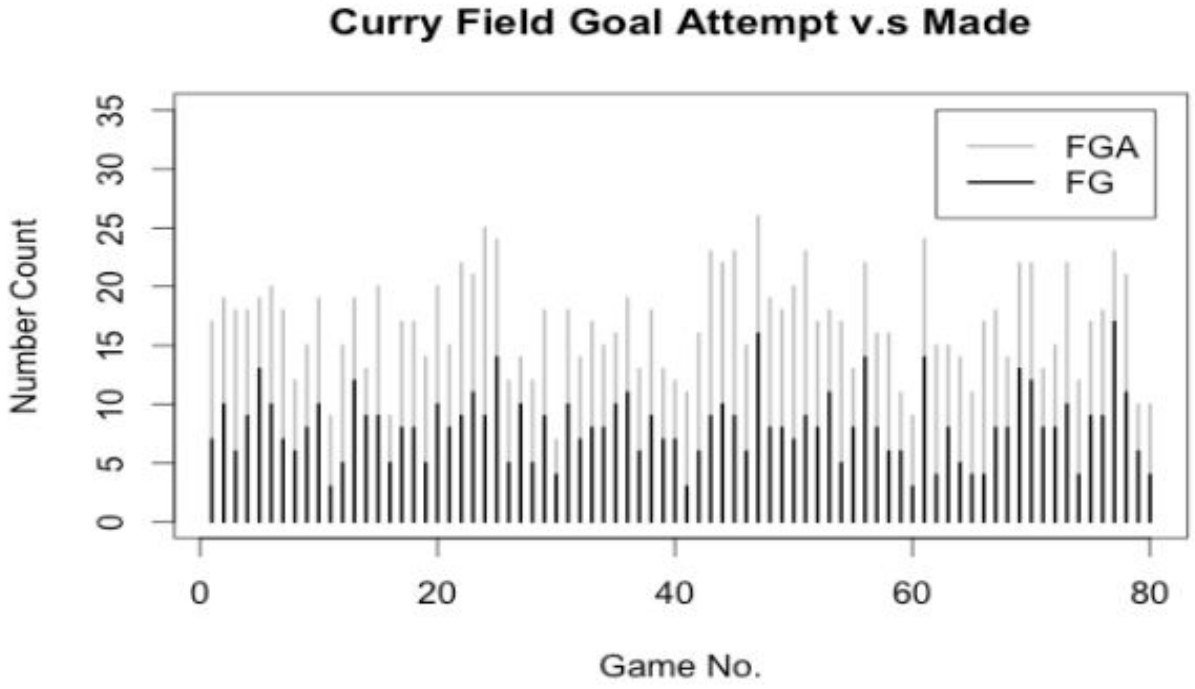

Histogram of curry$AST
Frequency
0
5
10
15
20
0
5
10
15
curry$AST

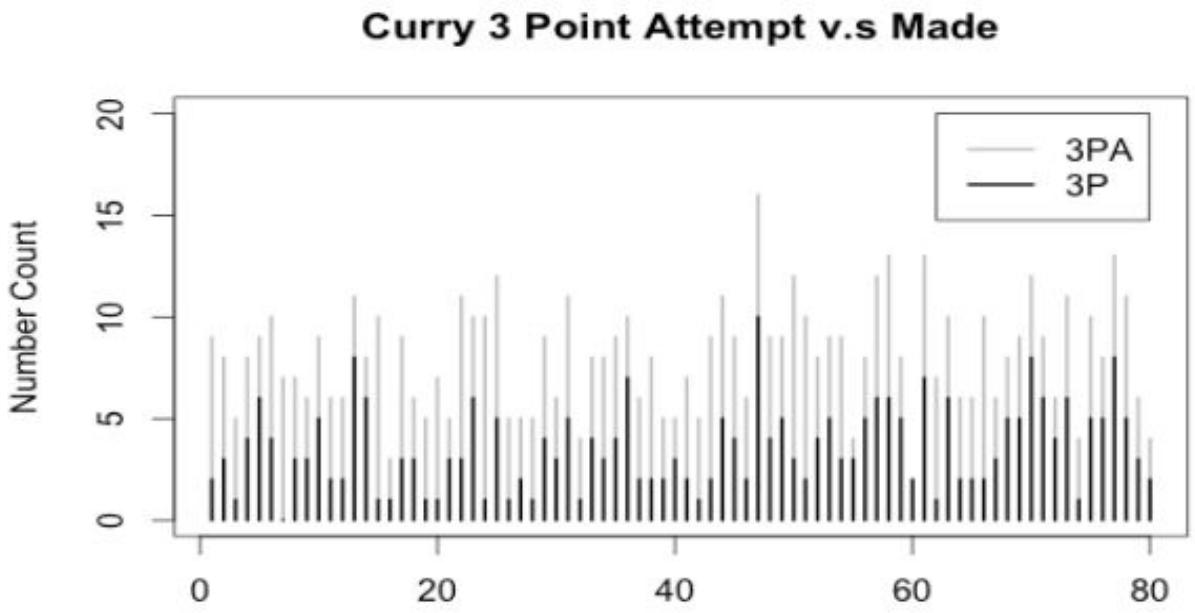

Curry$`AST`
Sample Quantiles
0
5
10
15
-2
-1
0
1
2
Theoretical Quantiles


Histogram of curry$TOV
Frequency
0
5
10
15
20
0
2
4
6
8
10
curry$TOV

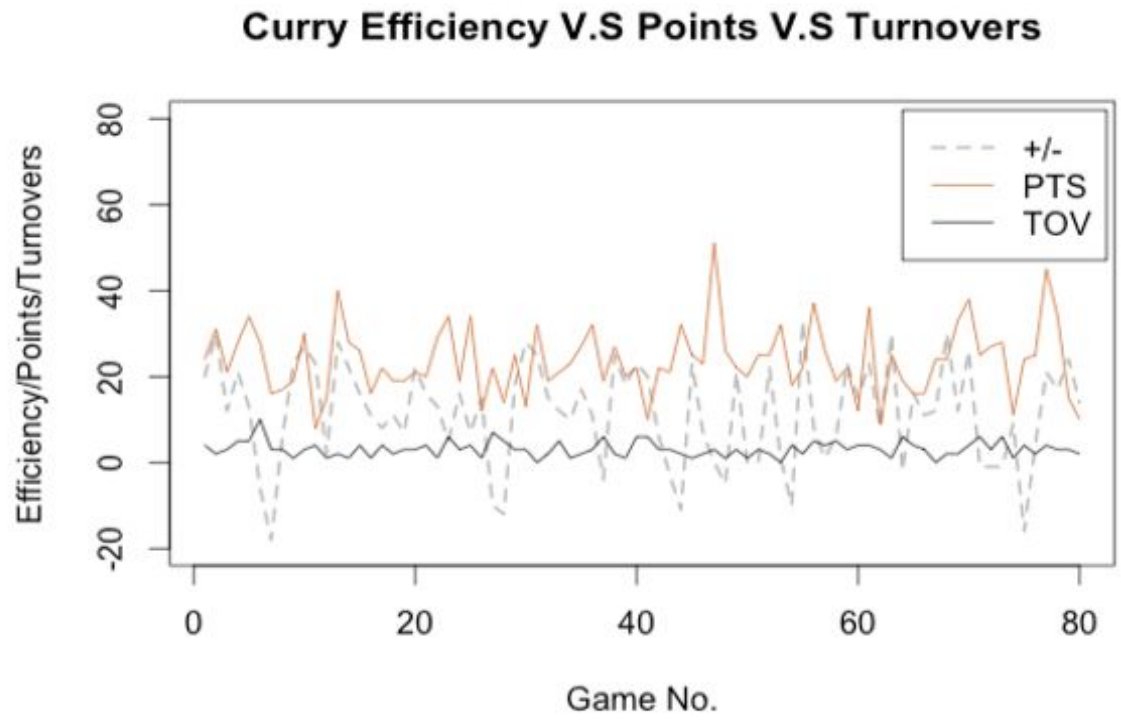

Curry$`TOV`
Sample Quantiles
0
2
4
6
8
10
-2
-1
0
1
2
Theoretical Quantiles

G2: Steph Curry's Field Goal Percentage & three point percentage over NBA 2014-2015 season.

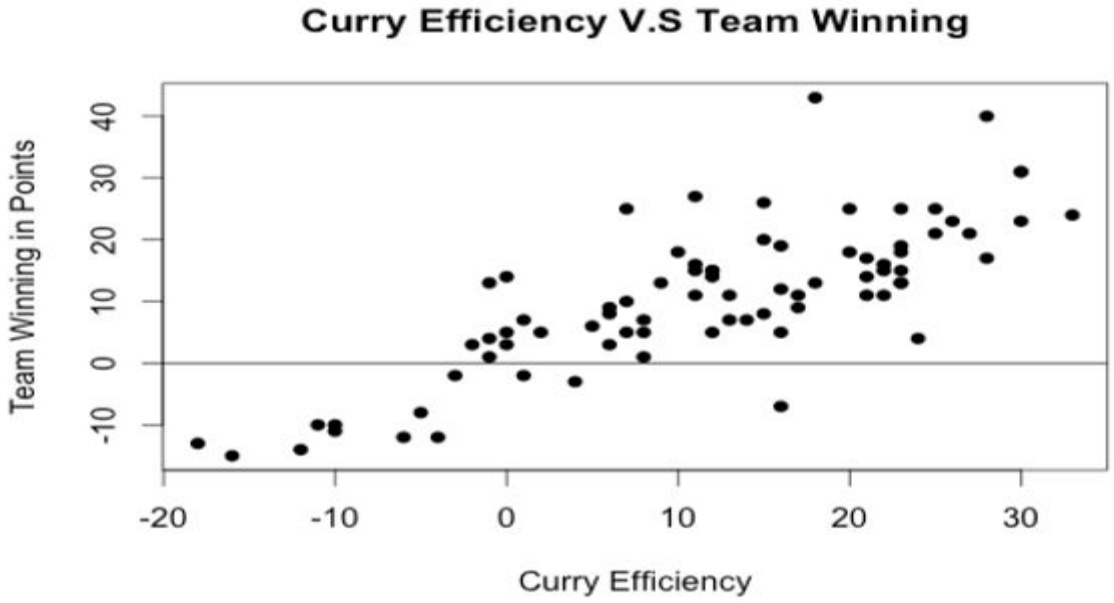


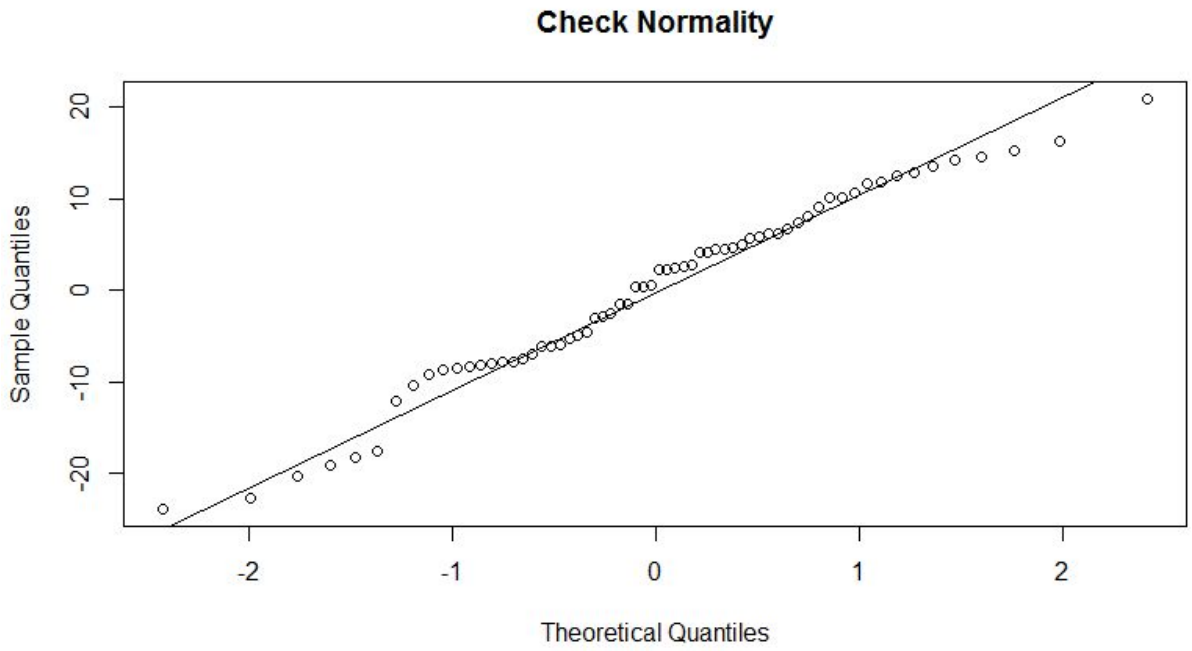
G3: Steph Curry's Efficiency/Total points per match/Total Turnover per match comparison over NBA 2014-2015 season



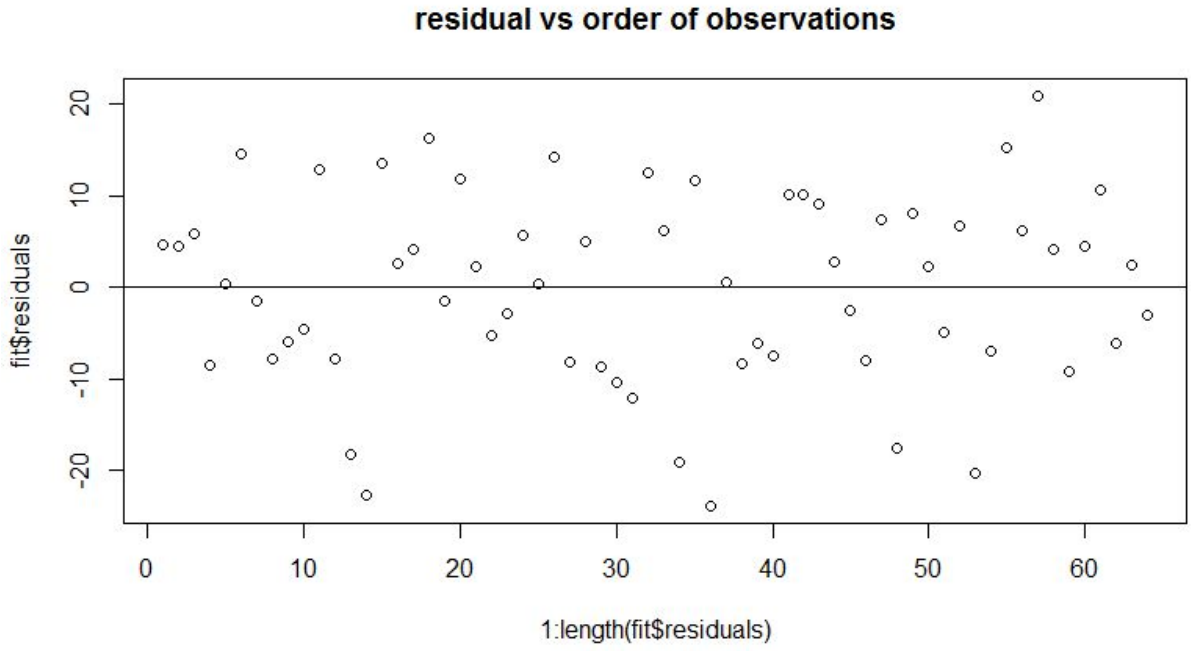

G4: Steph Curry's Efficiency V.S Team's winning



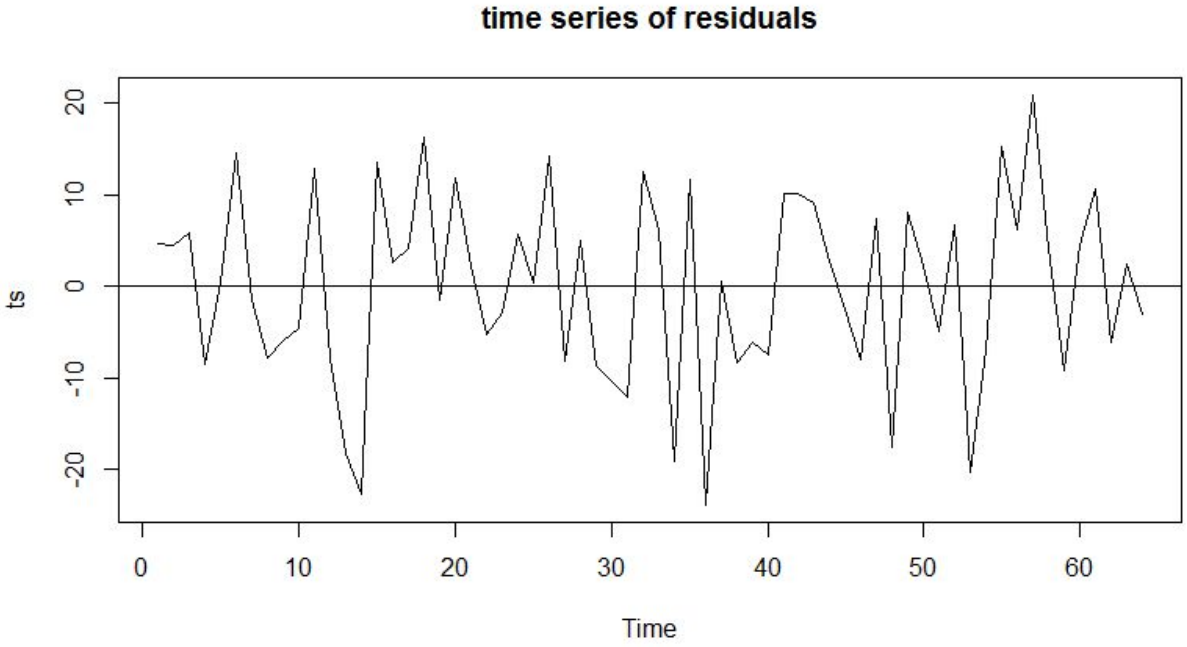
G5: qqplot of residuals




G6: residual vs order of observations



G7: Time series of residuals

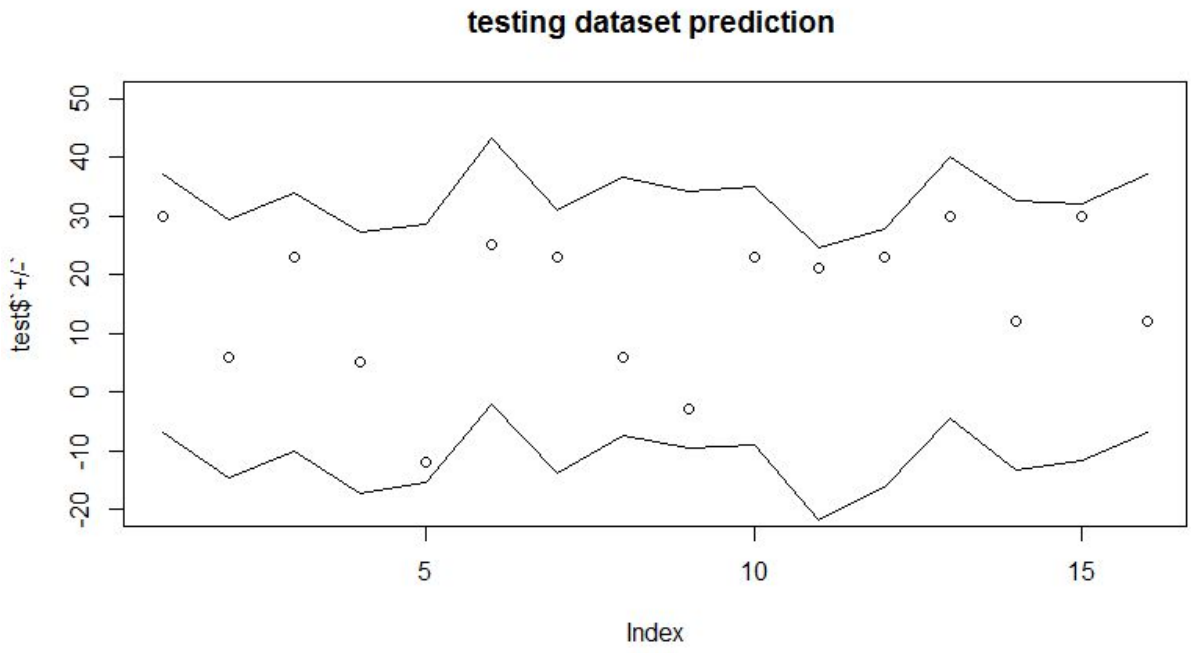



G8: Actual final MLR model output

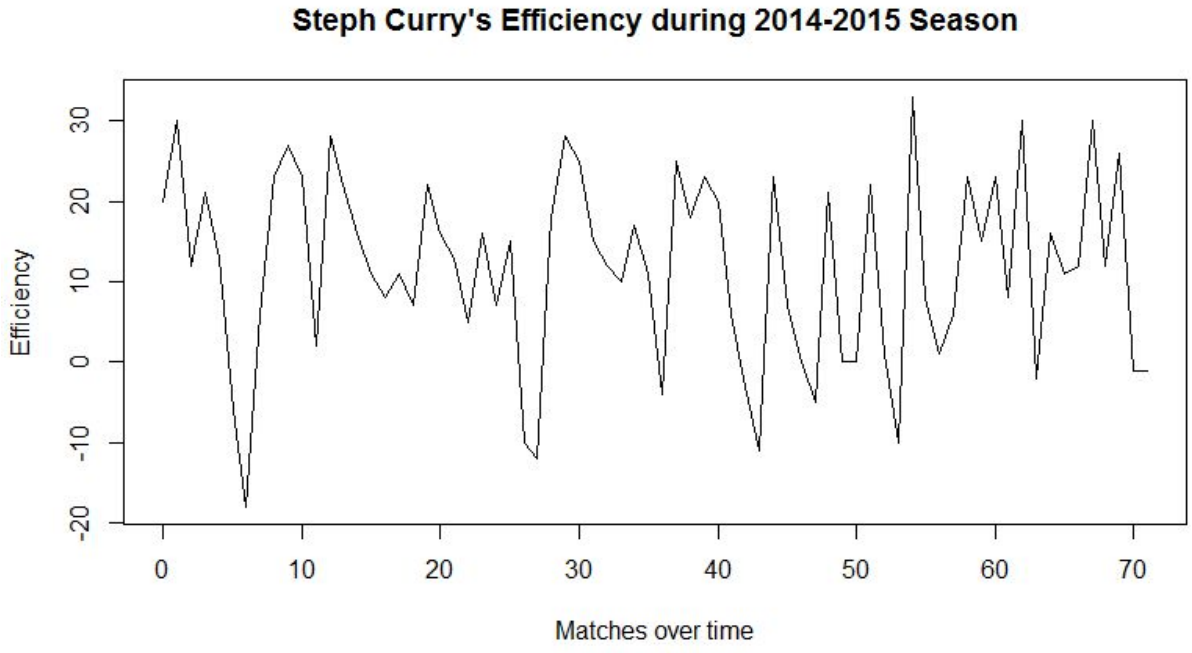
Final model: +/- = -6.5236 + 25.0483(FG%) + 1.3010(AST) - 1.6528(TOV)

```
Call:
lm(formula = train$`+/-` ~ train$`FG%` + train$AST + train$TOV)

Residuals:
    Min      1Q  Median      3Q     Max
-23.922  -7.563   1.404   6.841  20.919

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.5236     7.5222  -0.867   0.3893
train$`FG%`  25.0483    11.5528   2.168   0.0341 *
train$AST     1.3010     0.5028   2.588   0.0121 *
train$TOV    -1.6528     0.7368  -2.243   0.0286 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.51 on 60 degrees of freedom
Multiple R-squared:  0.2057,    Adjusted R-squared:  0.166
F-statistic: 5.178 on 3 and 60 DF,  p-value: 0.003005
```

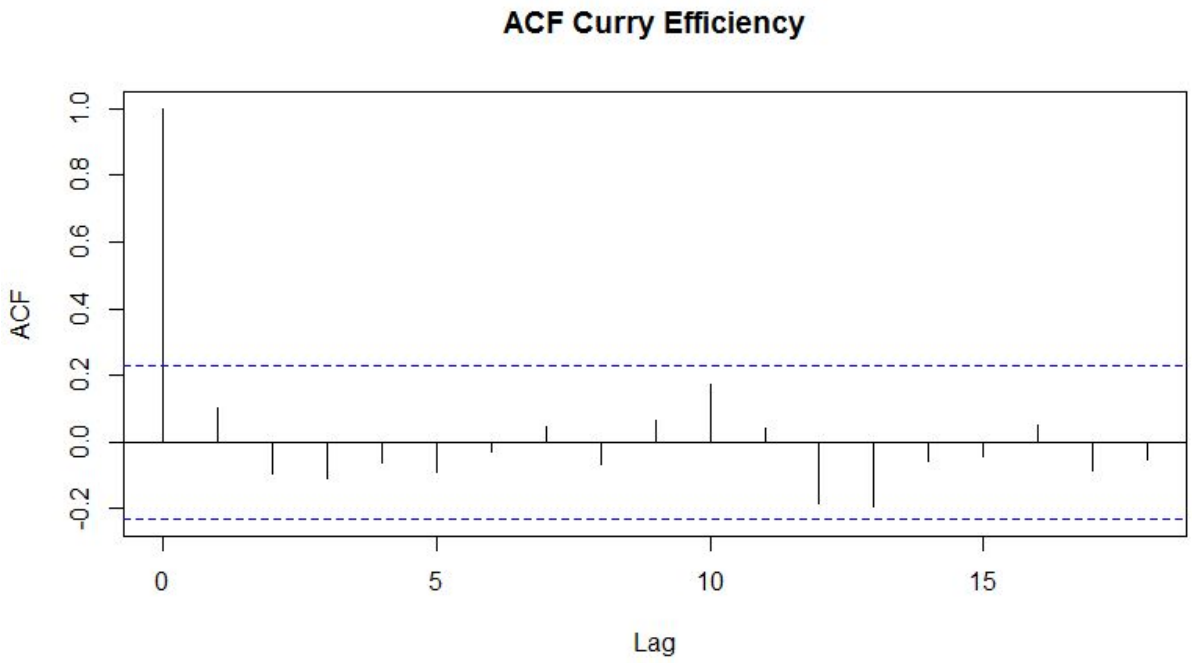
G9: prediction



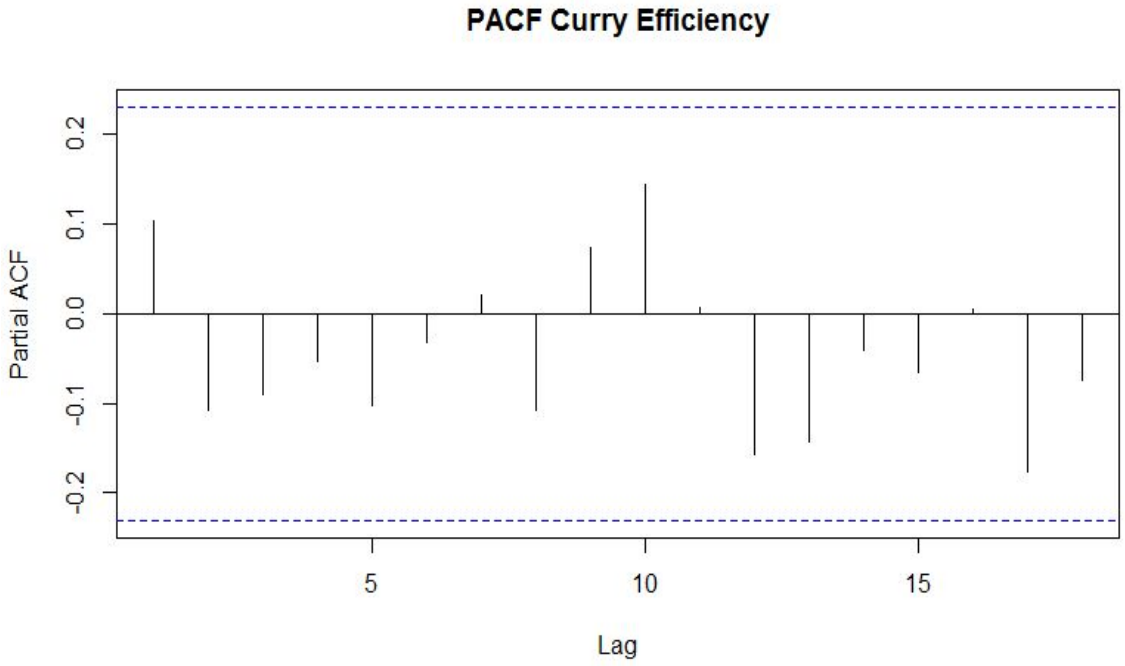

G10 : Curry’s efficiency Time Series plot



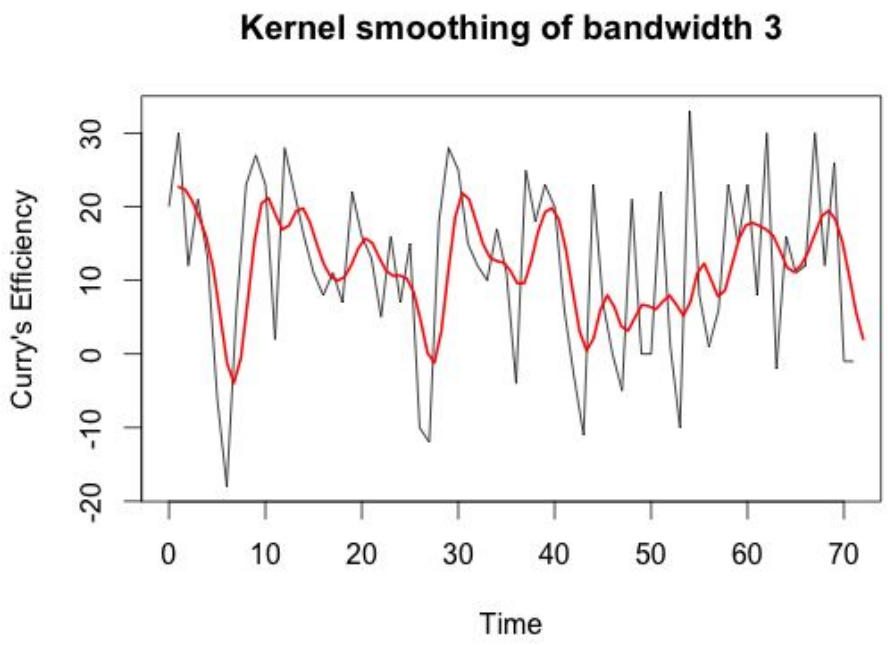
G11: ACF plot of Efficiency



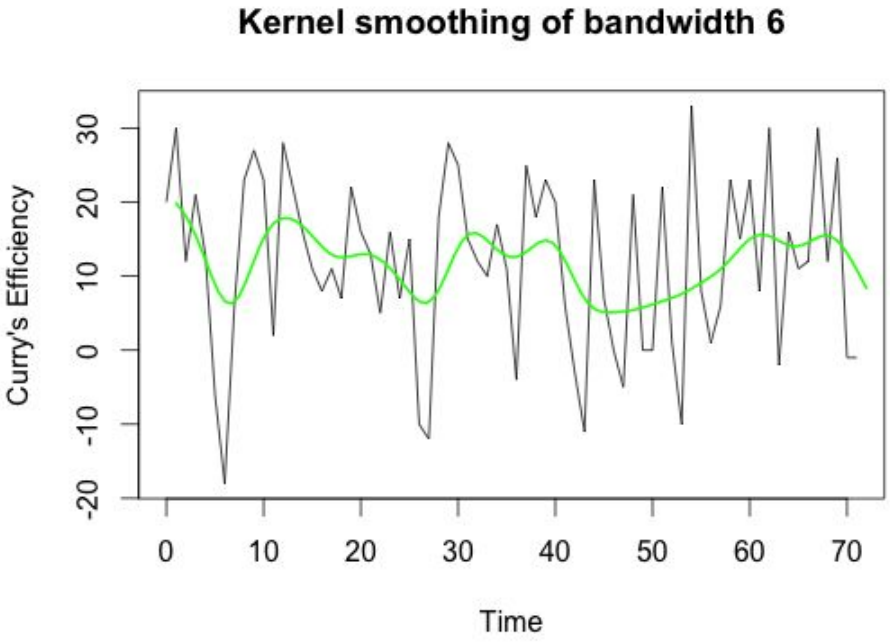

G12:PACF plot of Efficiency



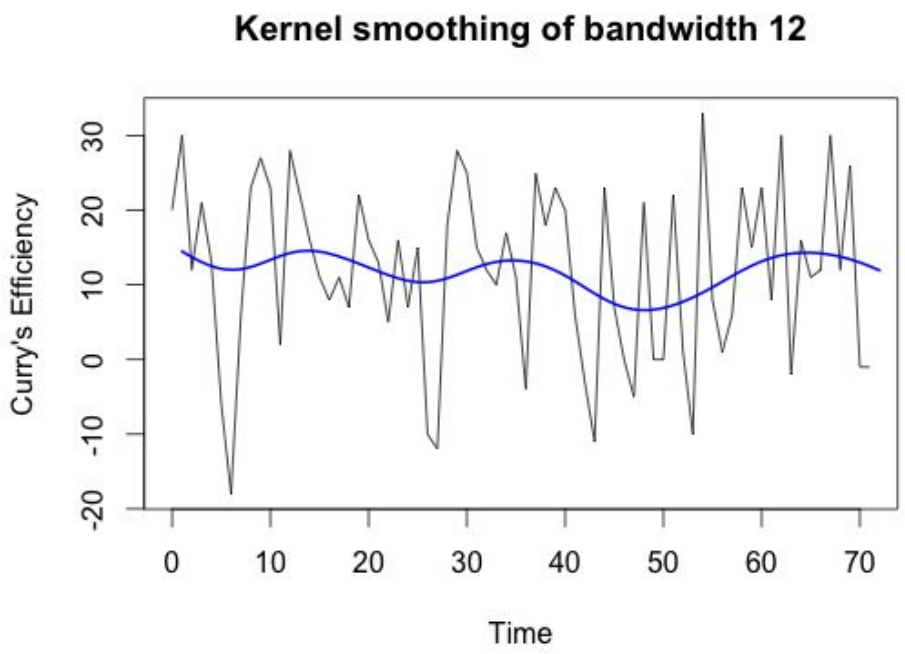
G12: Kernel Smoothing with Bandwidth 3



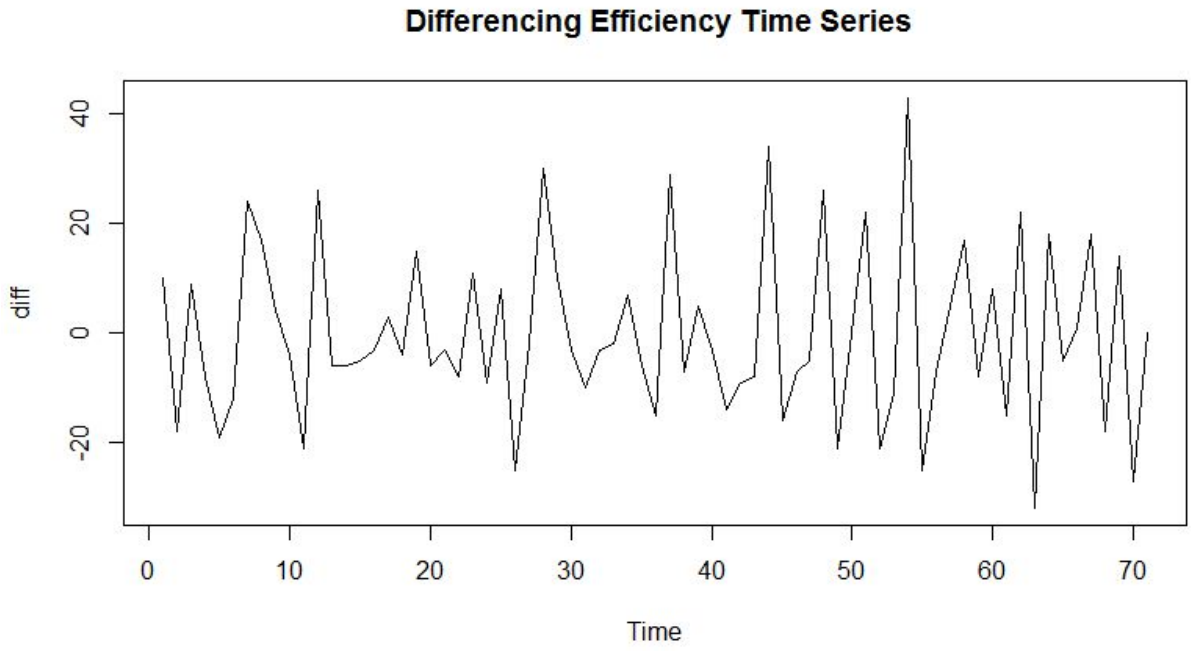

G13: Kernel Smoothing with Bandwidth 6



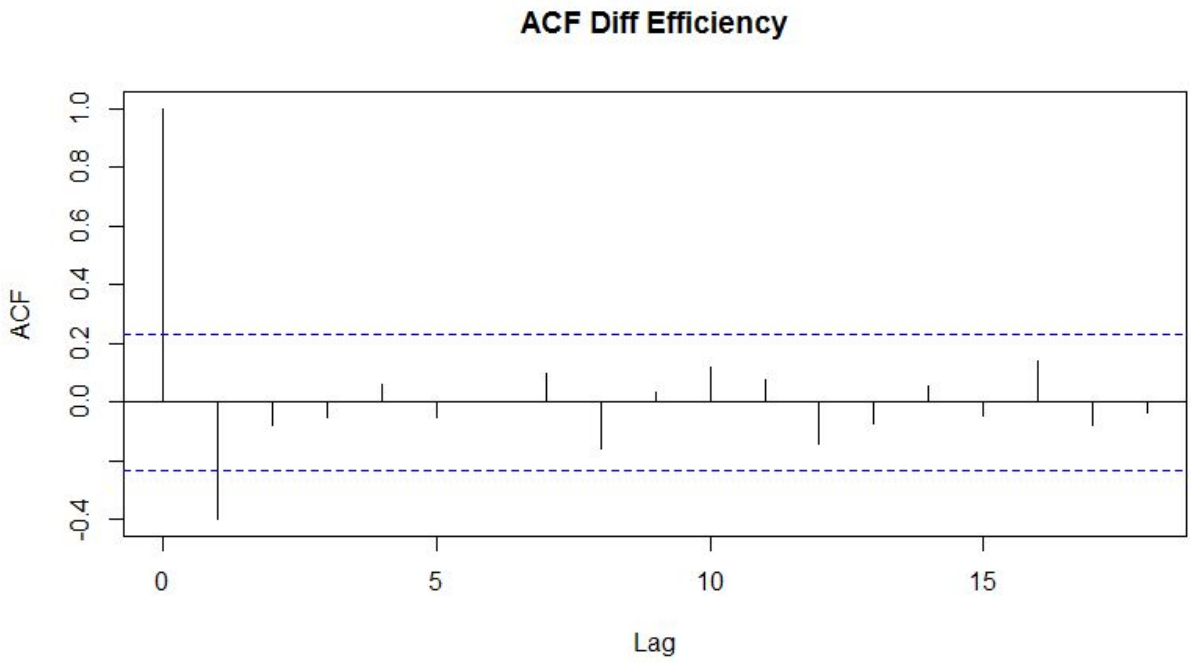
G14: Kernel Smoothing with Bandwidth 12



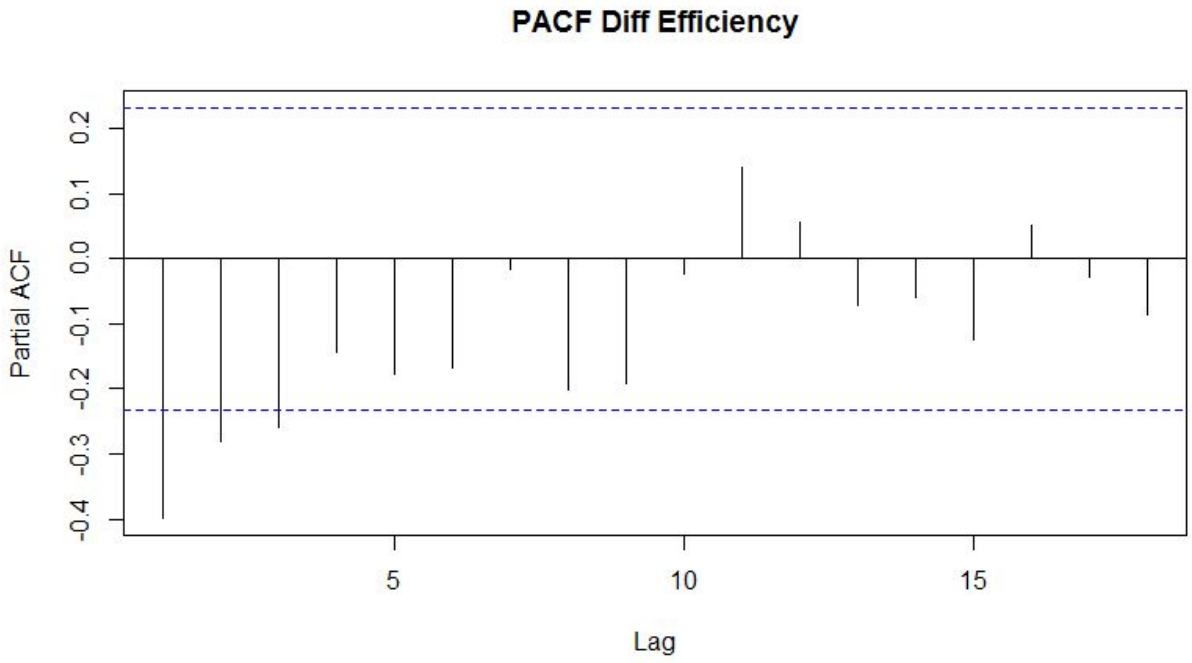

G15: Transformed Efficiency Time Series



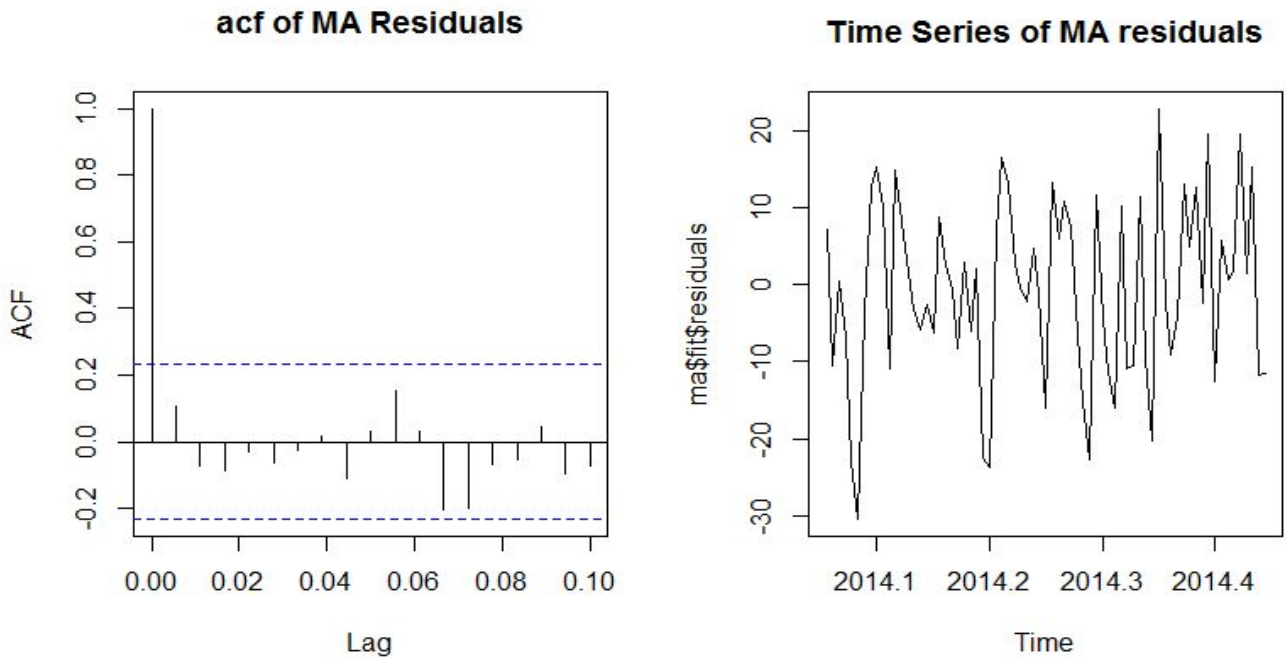
G16: ACF of Transformed Efficiency Time Series



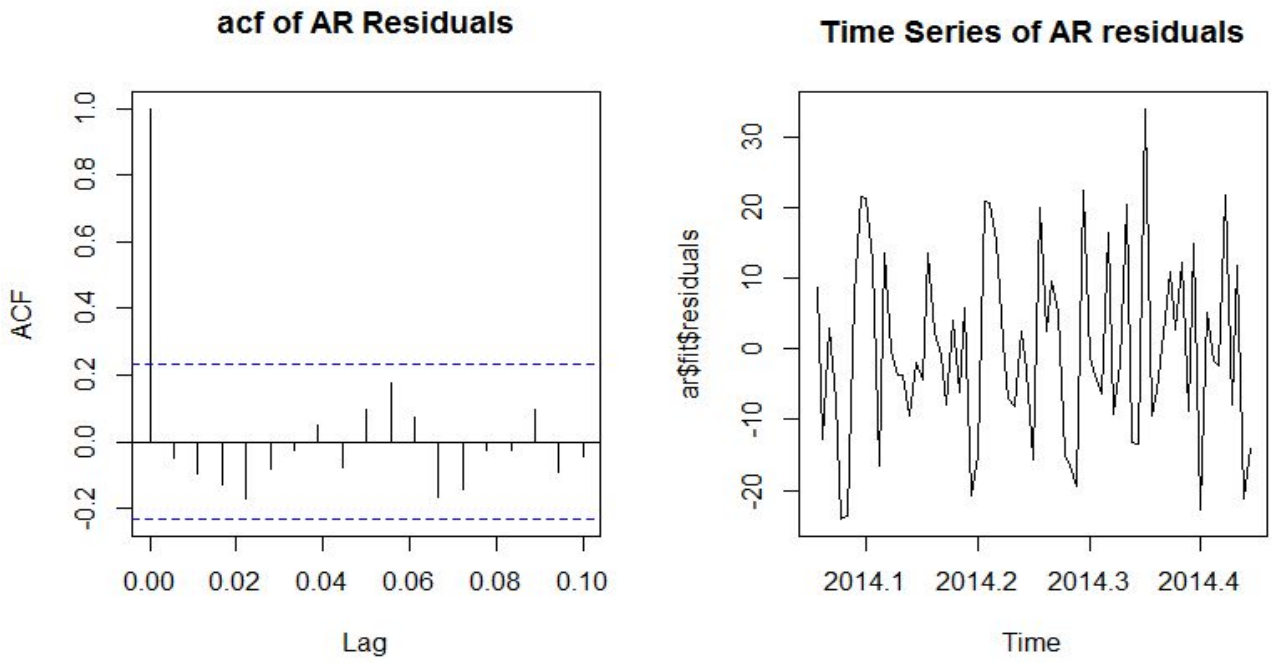

G17: PACF of Transformed Efficiency Time series



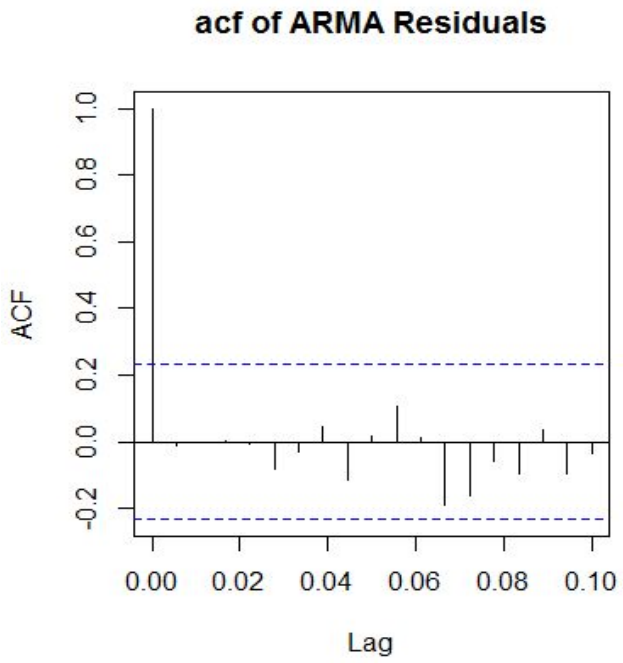
G18: Diagnostic of MA(1)

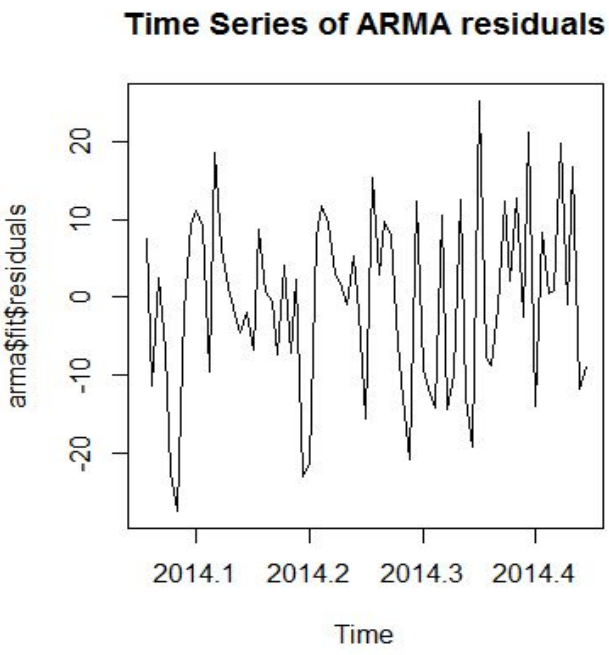



G19: Diagnostic of AR(3)



G20: Diagnostic of ARMA(3,1)




G21: Model Selection output

```
> c(ar$AIC,ma$AIC,arma$AIC)
[1] 6.281898 5.984034 6.037262
> c(ar$BIC,ma$BIC,arma$BIC)
[1] 5.409373 5.047772 5.196606
```

G22: Prediction using MA(1) model