**A Beautiful Dark Twisted Fantasy of Statistical Modeling:**

**An Attempt to Predict Three Point Shooting Percentage of NBA Teams**

Sabrina Martin, Jimmy McMillian

Period 7

Sports Stat

The following paper is about our process of creating a model to predict the three point shooting percentage of a given NBA team. The data used to create the model was from the 2014-2015 season, excluding playoffs. Our explanatory variables were assists, steals, and two-point percentage. The descriptive R-Squared for the 2014-20015 season was 61.9, however, when we ran the model against the 2013-2014 season data, the resulting R-Squared was -3.45. While our model seemed to work, it simply was not suitable for predictions in other seasons.

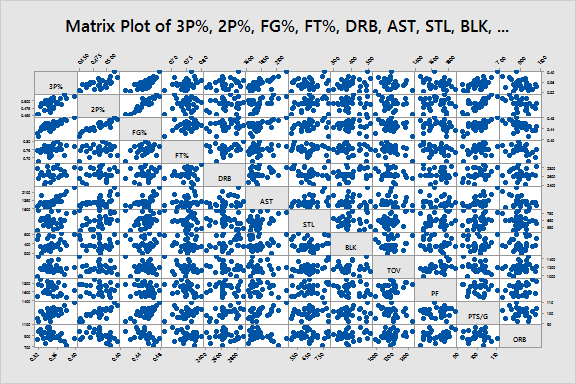
**Introduction**

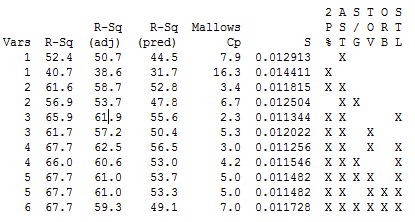
The NBA, North America’s most elite basketball league, boasts 30 teams and 82 games per season. When a team makes a basket, or a field goal, they score 2 points, or 3 points if it is made outside of the three-point arc. Free throws, awarded by 1 point, occur when the opposing team has a foul. An offensive rebound (ORB) is when the offensive team is able to catch the ball after it rebounds off the hoop, allowing the team to still keep the ball. A defensive rebound (DRB) is when defense obtains possession of the ball after the offense missed the shot. An is attributed to a player who passes the ball to a teammate in a way that leads to a score. A steal occurs when a defensive player causes a turnover. Blocks are when the opposing team prevents the offensive team from making the shot.

This paper will use different factors named “predictor values”, such as steals, blocks, etc. to predict a specific variable in each game.

**Methods**

To help determine which variables are correlated with one another, we used the matrix plot on Minitab to plot different variables against each other. In the matrix plot, we have graphs between 3-point percentage (3P%), 2-point percentage (2P%), field goal percentage (FG%), DRB, assists, steals, blocks, turnovers, personal fouls, points per game (PPG), and ORB. While 3P% and FG% seemed to have a linear pattern, indicating that they were correlated, this is because field goals are all 3-point and 2-point shots. It would also make sense that teams that are better at shooting 3 point shots, would be proficient 2 point shooting teams, because the difficulty of a shot increases with its distance from the hoop. Therefore a team that can shoot the more difficult 3-point shot well should be able to make 2-point field goals at a high rate as well. However, along with 3 point percentage and 2 point percentage, there seems to be a pattern with assists, steals, and offensive rebounds.

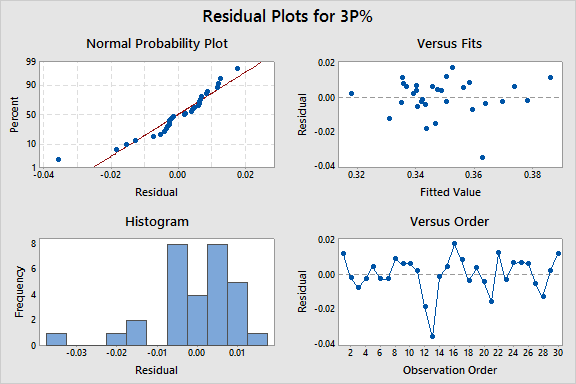




We then used the best subsets regression in Minitab to determine which combinations of predicting factors would best estimate the response variable. While more predictors would overall increase the correlation, too many is not worth the small increase in R-sq. Therefore, we wanted a model with a decently high R-Sq value, but no more than three predictor variables. Based on the R-sq values and the R-sq adjusted, we decided to use 2P%, AST, and STL as our predictors. 2P% would obviously correlate to the shooting ability of the team (though this does include close and “inside“ shots). AST would indicate the strength in the ball movement of a team which is an integral part of creating open looks, especially around the perimeter. Lastly, we believed that steals would indicate the strength of the guard play as well as give a better idea of the type of offensive opportunities that teams would receive. A team with more steals theoretically has more aggressive guards and a more high paced, full court offense. This type of play is positive for creating valuable and open scoring opportunities in the open court and creating more offensive possessions, and in turn scoring opportunities, in general.

**Results**

After using 2P%, AST. and STL as the predictors for 3P%, we created residual plots on Minitab. The equation for estimating **3P% = 0.0609 + 0.384 2P% + 0.000074 AST - 0.000051 STL**. For the 2014-2015 NBA season, using 2P%, AST, and STL as the predictors lead to a R-sq of 65.7 and a R-sq adjusted of 61.9.



After testing our regression equation with the 2013-2014 season, our R-sq was calculated to be -3.45, and our mean squared residual (MSR) was 0.00037. A negative R-Sq tells us that our model, based off of the 2014-2015 season was worse at predicting 3P% than simply using the mean. Though we are dealing with percents, a MSR in the tens of thousandths indicates that there was a relatively even distribution of results above and below our predictions.

**Discussion**

The residual plot is overall linear but becomes more curved towards the ends, meaning that the model we created is less accurate when the 3P% is at extremes. This also indicates that we could’ve use a quadratic model to fit this set better. The R-sq adjusted of 61.9 is acceptable, but not amazing. However, it does show some correlation between the predictor variables and 3P%.

When we tested our regression equation against another data set, the 2013-2014 season, our R-sq was calculated as a negative number of -3.45. This is much different from the R-Sq adjusted value of 61.9 from the previous data set, and it allows us to conclude that the predictor variables of 2P%. AST, and STL are unable to estimate well 3P%. Basketball and sports in general evolve over time. Last year in particular, the NBA saw an atomising riser in the shooting of 3-Point shots and an increase in overall flexibility of offenses. The model’s poor performance in the 2013-2014 season is just another showing of how the game changes over time, even on a year by year basis. While it seemed to work for the 2014-2015 season, it simply is not a suitable model for predictions in any other.