Forecasting NBA Player Performance using a Weibull-Gamma Statistical Timing Model

Douglas Hwang
Audible.com/Amazon.com
New York, NY, USA, 10001
Email: dh@douglashwang.com

Abstract

NBA franchises in many ways "risk" their future on certain players; this includes monetary commitments, salary cap restrictions, impact on overall franchise value, and team marketability. Instead of determining player value based on an individual players' past results (scoring average, eFG%, TS%, PER, Win Shares, +/-), it would be useful to determine player value by projecting future performance. This is especially important when examining long-term player contracts, and negating short-term bias in the statistically significant "contract" year performance. This paper explores player performance prediction based on a statistical timing model, by fitting a player's performance over time to a Weibull distribution, and accounting for unobserved heterogeneity by fitting the parameters of the Weibull distribution to a gamma distribution. By doing so, a statistical prediction of how a player might perform over the next few years can be determined, based on their trending performance from when they first entered the NBA. This will help predict performance over the next season, estimate contract value, and the potential "aging" effects of a certain player.

1 Introduction

To estimate an NBA players' performance is often similar to picking roulette numbers at a casino. The number of variables involved to get a specific understanding of their performance, are too many to understand. How can one accurately determine a players' own innate ability? How is learning incorporated? How is coaching style incorporated? How are different team members factored?

It has been noted that the free agent class of 2010 was one of the best ever. Not only including marquee names like Lebron James and Dwayne Wade, but a host of experienced "superstars." It is important have an understanding of their past contributions on court, but the key question is, what is the expected performance in the contact years to come?

This paper primarily focuses on points (though the model can be applied to other metrics as well) and how point production can be forecasted in a specific player's future years.

2 Background

To adequately understand a player's value on the basketball court, beyond basic metrics (points, assists, rebounds, etc…), currently there are a number of "advanced" metrics that are being employed by various NBA teams. A few examples are below (the detailed formulas can be seen in the appendix):

1. **NBA Efficiency Formula** - This is a commonly gauge used when evaluating a players' contribution to a team [1]. It uses commonly used, simple to understand, takes easy to obtain stats and combines it into one metric, and often used in free agent salary determinations. The deficiency with this metric for salary determination is that it is based on historical achievements, rather than future expected contributions.

2. **Player Efficiency Rating (PER) -** This is a formula developed by ESPNs John Hollinger that determines a per minute rating of the player [2]. This calculation makes an unadjusted PER, which is then scaled so that the average player for a specific season is 15. Higher PER means the player is better relative to others in the league.

3. **Win-Shares** – This is a measurement that determines how many wins to a team that was contributed by a certain player. Modeled after Bill James' baseball win-shares, it was developed by Jason Kubatko while working for basketballreference.com [3].

4. **Plus-Minus/Adjusted** – This is a system that was modeled after hockey, where it determines how points the team had while the player was on the court. Though the NBA tracks the absolute plus-minus, most teams/references use an adjusted plus-minus system, which is specifically weighted to each teams' own metrics [4]. There is no generally agreed upon system in place yet.

The underlying story for these metrics is that a player's contribution on the court depends on a variety of factors and weights. Each contribution may have a different weighting based on context (team dynamics, environment, competition).

3 Methodology

Originally, after trying a multitude of combinations of the "all-encompassing" formulas to predict future performance, it was decided to try a much more generic, simple, and commonly used statistic: points. This singular number, points, is determined by the "system" (team-mates, teams, opponents, health, etc…), but in order to gain points, the player himself will need to score. Hence, still a measure of how a player works within their environment. The following developed model was able to forecast points for a particular player within a FG.

In forecasting a player's ability to score, it had to be normalized for the number of games they played. This alone doesn't tell the full story, since often some years are not like the other "normal" years. For instance, a player can have an injury that affects them the whole season, decreasing their contribution. On the flip side, a player could be in a contract year or just particularly motivated, leading to abnormal contributions for that year. Empirical observations have shown that contract years tend to lead to higher numbers [5].

Using points, a statistical model was used to predict the future performance of that player. Akin to estimate a students' performance, based on a normal distribution (variance and mean), a similar statistical model was used to estimate point performance. This concept was first inspired by paper on basketball free throw percentage using Bayesian statistics by Richey and Zorn, in Mathematics

Magazine [6]. Then further applied by paper on NFL field kickers luck/skill by using a negative binomial distribution by Morrizon and Kalwani [7].

Using a Weibull hazard rate function, mixing with a gamma function, and including co-variates, a Weibull-gamma with covariates model was created as seen below (detailed math seen in Appendix),

$$P(T \leq t) = \int_0^\infty \left(1 - e^{-\lambda \cdot D(t)}\right) \cdot \left(\frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)}\right) d\lambda$$
$$= 1 - \left(\frac{\alpha}{\alpha + D(t)}\right)^r$$

To solve for the parameters of this function, the model used a method of maximizing log likelihood. And for this case, only a single covariate was used to explain an "off-year," normal year, or an "on-year." An off-year, most often an injury year, but can also take into account major coaching/benching roles. An "on-year" is often a contract year or some other motivated reason to have an uncharacteristic season.

4 Results

First to test the model, statistics were obtained for previous players who had already finished their career, and the modeled was applied to it to see how it stacked.

In this example, Charles Barkely's career is seen. Given that his career was done, I matched the model the entire career (figure on left). Once that was obtained, a holdout sample of 7 years was used to determine the accuracy of the model's ability to match the rest of the player's career (figure on right). Both the calibrated and holdout sample mean absolute percentage error (MAPE) is shown.

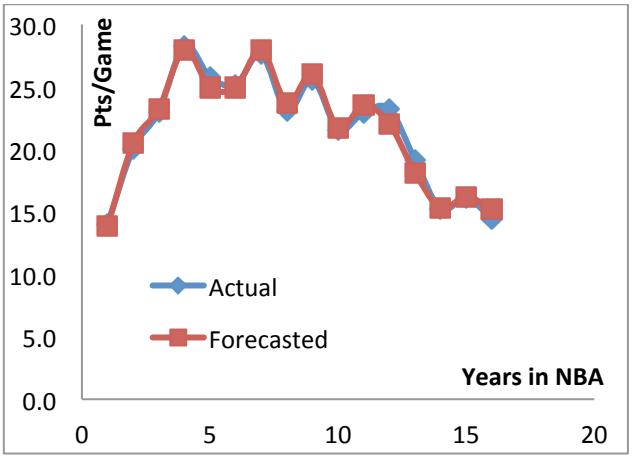


Figure 1: Charles Barkely career model matching. Calibrated MAPE - 0.66%

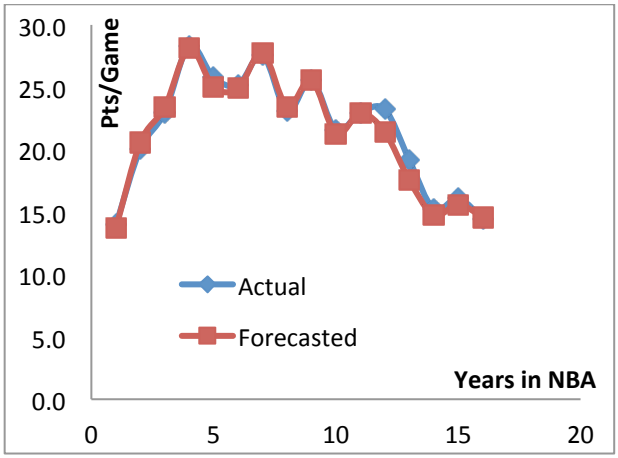


Figure 2: Charles Barkely career projections after 7 years. Holdout MAPE - 0.79%

Other test "legends" are shown in the appendix. After obtaining verification with the test players, this model was applied to the top free agents in the summer of 2010, and verifying the results to their 2011 numbers. According to nbafanhouse.com [8], the top 8 rated free agents in the 2010 summer were (in order): Lebron James (7 years), Dwyane Wade (7 years), Chris Bosh (7 Years), Dirk Nowitzki (12 years), Yao Ming (8 years), Joe Johnson (9 years), Amare Stoudemire (8 years), and Carlos Boozer (8 years).

Specifically looking in detail of two of the aforementioned 2010 summer free agents, the projections can be seen below.

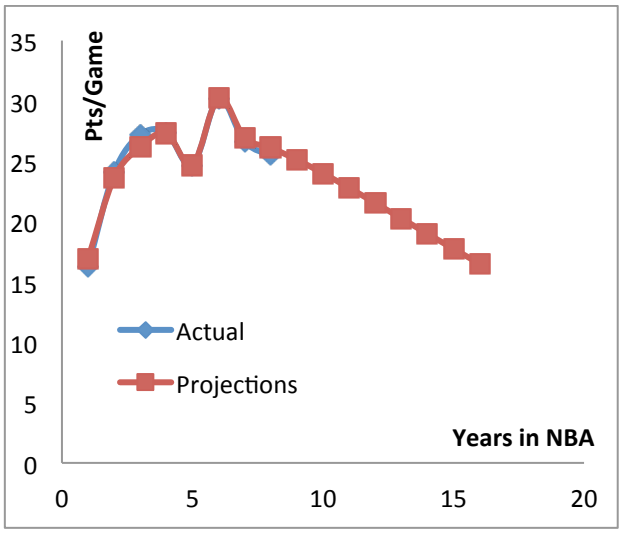


Figure 3: Dwayne Wade Projections

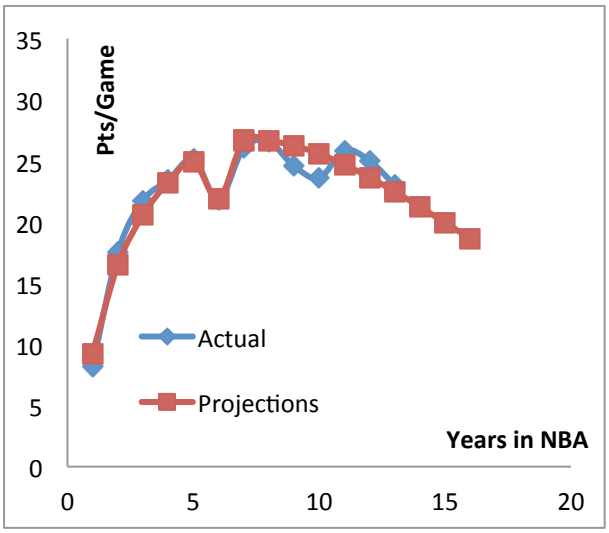


Figure 4: Dirk Nowitzski Projections

Using the data from 2010-11 season and back, the projections are shown below. The free agents of the 2010 summer class are shown with their actual 2010-2011 season number, the model predictions, and the differences between the two.

Table 1: Projections vs Actual for 2010-2011 season for seven 2010 summer free agents

	Actual Pts/Game 10-11 Season	Pts/Game Model Prediction	Difference
Lebron James	26.7	26.2	-0.5
Dwyane Wade	25.5	24.1	-1.4
Chris Bosh	18.7	20.2	1.5
Dirk Nowitzki	23.0	22.4	-0.6
Joe Johnson	18.2	20.1	1.9
Amare Stoudemire	25.3	26.5	1.2
Carlos Boozer	17.5	16.4	-1.1

5 Discussion

The reason a Weibull-Gamma (WG) model was used vs other models, was that a timing model was needed. The key parameter of interest is performance with respect to time. Other models such as an exponential-gamma would not fulfill the underlying story, as basketball performance is not "memory-less." Time has a strong effect on athletic ability, and in turn basketball performance. "Learning" also plays a role in a players' development and this model takes that into account. The WG model allows for unobserved heterogeneity, and uses the underlying hazard rate to model the individual timing behavior/performance.

Another model in consideration, similar to NFL field kickers by Morrizon and Kalwani [7], was a binomial model. But again, it would not fit the underlying story, as a binomial is a "yes/no" decision. In the case of scoring, every player scores an unknown amount of points based on their own innate ability and environment.

The test subjects in this model (the "legends") portray an accurate connection between the model and their own careers. This gives confidence to move forward with current players. The rationale for the binary covariate use, of the "hot" "cold" years is to leave the model relatively simple. By trying to fit whether the player has a "more" "hot" year or not goes beyond the scope of the model underlying principle, of keeping it simple.

When comparing this season's results with the 7 free agents of 2010, and their actual performance, again the data fit the model well, all within one field goal per game. The contract year extrapolation was interesting for the listed players, as NBA GMs must've done comparable evaluations to determine player contracts.

The core concept in all the predictions is the hazard rate function represented by the figure below.

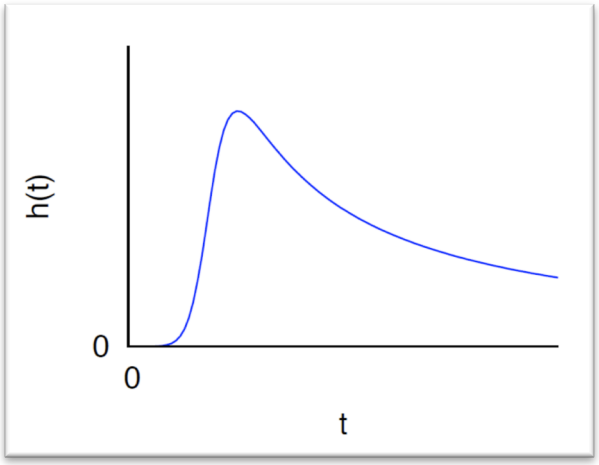


Figure 5: Hazard rate function graph

This underlying hazard rate model makes conceptual make sense when applied to scoring rates of basketball players. That is, as a basketball player enters the NBA, they would first "learn the ropes" become familiar with the league and style. As they develop their athletic ability and become more familiar, every player hits a crest and peaks in performance. Then, as time overcomes all, athletic ability slowly diminishes, likely decreasing the ability for the player to score.

6 Conclusions

Using probability models to forecast the performance of NBA players is an excellent way to evaluate the future "portfolio assets" for the team, franchise, and business. Investing in NBA players is expensive, and instead of determining value from the "gut," or from what the "market" thinks, the value should be determined by using a combination of quantitative analysis with the right qualitative analysis.

If this analysis can be done with points, it can be done with assists, rebounds, turnovers, and all basic basketball metrics. The best utility of these models is not to make an analysis and let the analysis sit for the next contract season, but as new information arrives, to update the model to stay current on player performance trends and monitor any/all warning signs. As each year passes, it would be best to update the model and see where the new trajectory is pointing. Since the model shows the most statistical likely outcome with the information on hand to date, as more information is obtained, a better understanding of the underlying story can be made, and a slightly better expectation of the future can be gleaned.

Though this insight gain to most may be minor, as learned in Moneyball any/all additional insight relative to the competition can mean the difference between lottery, playoffs, and the Larry O'Brien trophy.

7 Acknowledgements

Thanks to my family and friends who have supported me throughout life and random projects and endeavors (Hosuk Hwang, Eunsoo Hwang, Jennifer Hwang, and Michelle Luk).

Thanks to Clarence Lee, Ph.D candidate at Harvard Business School, for making sure I didn't make things up in this paper.

Thanks to Prof. Michael Braun, who showed me the powerful application of Bayesian statistics in the Customer Probability models class at MIT Sloan.

8 References

[1] http://www.nbastuff.com
[2] http://en.wikipedia.org/wiki/Player_Efficiency_Rating
[3] http://www.basketball-reference.com/about/ws.html
[4] http://www.82games.com/ilardi1.htm

[5] Stiroh, Kevin. "Playing for Keeps: Pay and Performance in the NBA". Economic Inquiry. Jan 2007.
[6] M. Richey and P. Zorn, "Basketball, Beta, and Bayes," Mathematics Magazine, vol. 78, no.5, pp. 354-367, Nov. 2005.
[7] D.G. Morrison and M. U. Kalwani, "The Best NFL Field Goal Kickers: Are they Lucky or Good?" Chance: New Directions for Statistics and Computing, vol. 6, no. 3, 1993.
[8] http://nba.fanhouse.com/2010/04/06/top-50-2010-nba-free-agents/

9 Appendices

NBA Advanced Metric Formulas

1. **NBA Efficiency Formula** - (Points)+(Rebounds)+(Steals)+(Assists)+(Blocked Shots)-(Turnovers)-(Missed Shots)

2. **Player Efficiency Rating (PER)**

```
uPER = (1 / MP) *
    [ 3P
    + (2/3) * AST
    + (2 - factor * (team_AST / team_FG)) * FG
    + (FT *0.5 * (1 + (1 - (team_AST / team_FG)) + (2/3) * (team_AST / team_FG)))
    - VOP * TOV
    - VOP * DRB% * (FGA - FG)
    - VOP * 0.44 * (0.44 + (0.56 * DRB%)) * (FTA - FT)
    + VOP * (1 - DRB%) * (TRB - ORB)
    + VOP * DRB% * ORB
    + VOP * STL
    + VOP * DRB% * BLK
    - PF * ((lg_FT / lg_PF) - 0.44 * (lg_FTA / lg_PF) * VOP) ]

  Where,
  factor = (2 / 3) - (0.5 * (lg_AST / lg_FG)) / (2 * (lg_FG / lg_FT))
  VOP    = lg_PTS / (lg_FGA - lg_ORB + lg_TOV + 0.44 * lg_FTA)
  DRB%   = (lg_TRB - lg_ORB) / lg_TRB
```

3. **Win-Shares** –

   a. **Calculate points produced for each player.** In 2008-09, James had an estimated 2345.9 points produced.
   b. **Calculate offensive possessions for each player.** James had an estimated 1928.1 offensive possessions in 2008-09.

c. **Calculate marginal offense for each player.** Marginal offense is equal to (points produced) - 0.92 * (league points per possession) * (offensive possessions). For James this is 2345.9 - 0.92 * 1.083 * 1928.1 = 424.8. Note that this formula may produce a negative result for some players.
d. **Calculate marginal points per win.** Marginal points per win reduces to 0.32 * (league points per game) * ((team pace) / (league pace)). For the 2008-09 Cavaliers this is 0.32 * 100.0 * (88.7 / 91.7) = 30.95.
e. **Credit Offensive Win Shares to the players.** Offensive Win Shares are credited using the following formula: (marginal offense) / (marginal points per win). James gets credit for 424.8 / 30.95 = 13.73 Offensive Win Shares.

4. **Plus-Minus/Adjusted** – There is no completely agreed upon system in place yet.

Weibull-Gamma with covariates Model

Using a Weibull with co-variates hazard rate function:

$$F(t) = 1 - e^{-\lambda \cdot D(t)},$$

where,

$$D(t) = \sum_{i=1}^{t} [i^c - (i-1)^c] \cdot e^{(\beta' x(i))}$$

and mixing with a gamma function,

$$g(\lambda) = \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)}$$

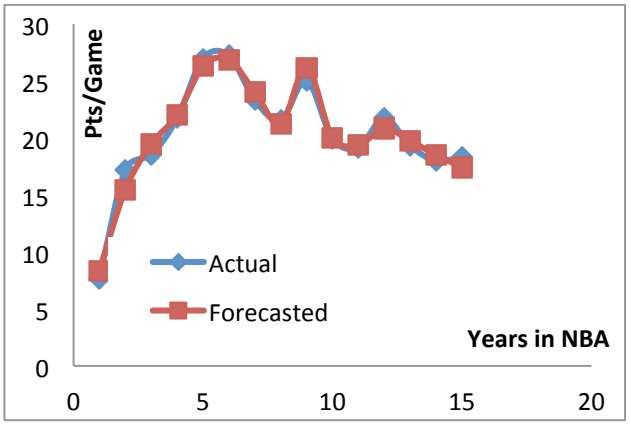
gives

$$P(T \le t) = \int_0^\infty \left(1 - e^{-\lambda \cdot D(t)}\right) \cdot \left(\frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)}\right) d\lambda$$

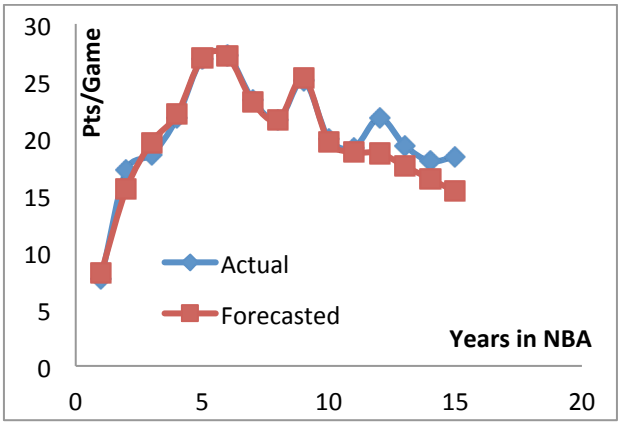
$$= 1 - \left(\frac{\alpha}{\alpha + D(t)}\right)^r.$$

Weibull-Gamma with covariates Model applied to Legends:

Clyde Drexler:

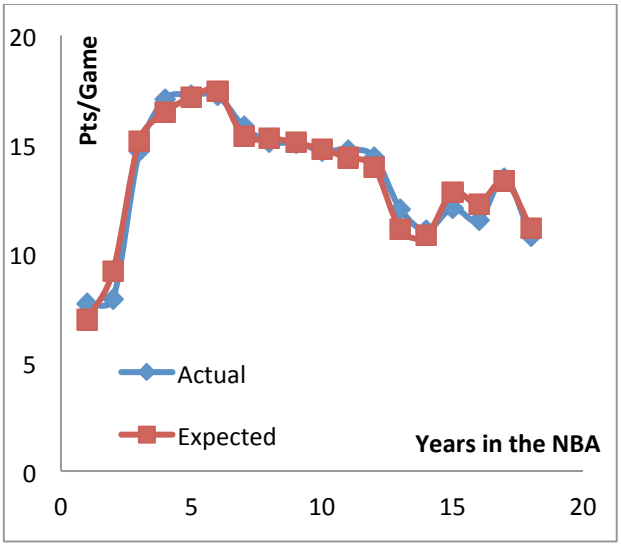


Calibration MAPE - 1.56%

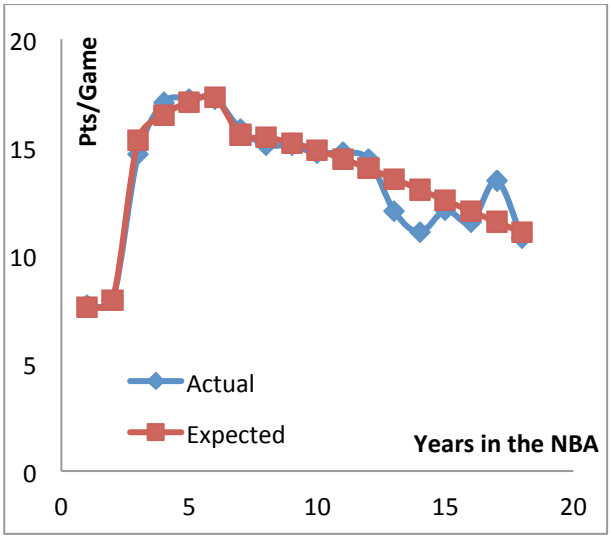


Holdout MAPE - 1.34%

John Stockton

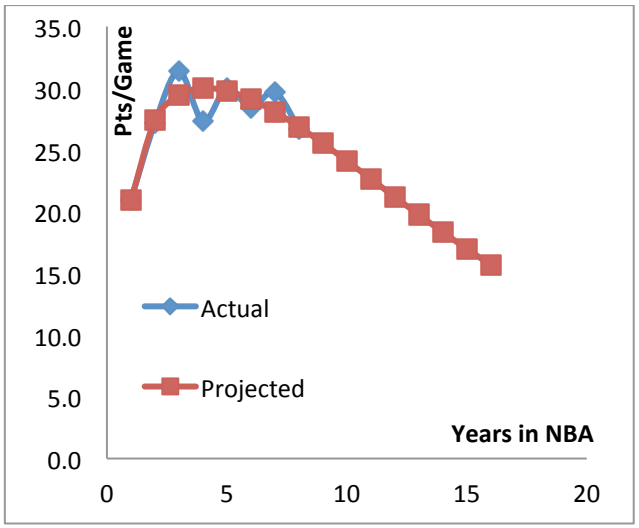


Calibration MAPE - 0.52%

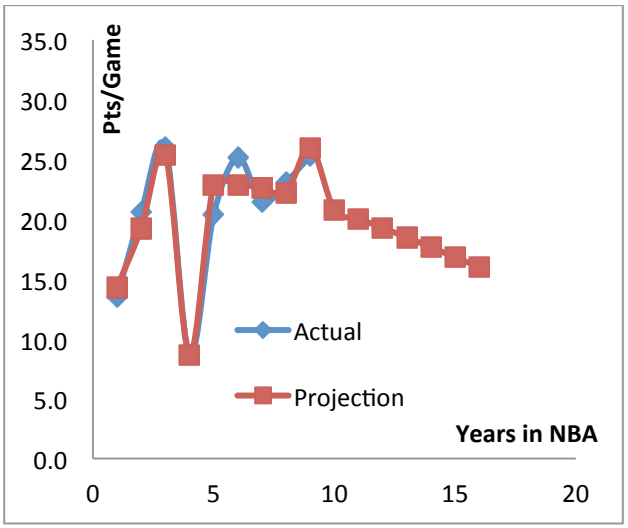


Holdout MAPE – 0.81%

Weibull-Gamma with covariates Model applied to other 2010 free agents:



Lebron James Projections vs Actual



Amare Stoudemire Projections vs Actual