Projecting Fantasy Football Points

Brian Becker
Gary Ramirez
Carlos Zambrano

MATH 503 A/B

March 15, 2018

1 Abstract

Fantasy Football has been increasing in popularity throughout the years and becoming a popular way for participants to earn money. Participants form leagues in which they draft NFL players into their teams to compete against other teams in their leagues. Participants utilize projections of players' overall Fantasy Points from various websites to help them draft the highest scoring players to their team in the beginning of each season. Our goal is to reproduce and improve these projections for the quarterback position. We will use saved data from the previous 2008 - 2014 NFL seasons to train various linear regression models. We will compare the results of our various linear regression models against one other while also striving to surpass Fantasy Football Analytics projections performance.

2 Introduction

Fantasy Football is a growing industry where participants go head-to-head on a weekly basis to see which of their teams can get the most points. These points are accumulated based on player performance, so picking the "best players" can take one anywhere from several minutes to hours. Teams are initially picked in Fantasy Drafts, which consist of 6-14 participants. Participants take turns in selecting players for their team; however, there are several ways in which the teams are selected. For example: some participants go after their favorite players; others quickly select the players with the highest amount of projected fantasy points. These projected points are provided by several different experts within the fantasy football world, but how does one come up with such projections?

Fantasy football points are known to be difficult to project. For example, even a well-regarded player who regularly scores his projected points might be pulled out of a game due to injury or suspension. This leads to a very high variation in each player's fantasy points. This is the main difficulty in projecting fantasy points for each player; players do not typically score similar fantasy points in consecutive seasons. While utilizing a player's past performance, we would also like to find a way to account for a player's of not playing the same number of games or not obtaining enough ball *touches* (where the player could subsequently score fantasy points).

3 Literature Review

There are many Fantasy Football websites out on the web that offer"black box" type projections. Some of the most commonly used websites offer fantasy football projections include ESPN, FantasyPros, FantasySharks, NFL.com, and Yahoo. In fact, Isaac Peterson [10] from *Fantasy Football Analytics* analyzes the accuracy of these types of projections of seventeen of such websites; for the aforementioned five websites' projections he computes the *coefficient of determination*, R^2 and the *mean absolute scaled error*, $MASE$ statistics to be:

Source	R^2	$MASE$
ESPN	.483	.591
FantasyPros	.547	.516
FantasySharks	.455	.547
NFL.com	.474	.612
Yahoo	.499	.567

Table 1: 2014 Projection Evaluations by FFA

We can think of R^2 as a measure of how much of the variance of the dependent variable is explained by the model [6], and the $MASE$ statistic as a relatively new, scale-free measure of the accuracy of forecast models [5]. A good model will ideally capture a high R^2 and a low $MASE$. With such low R^2 values, we can see that these websites struggle forming accurate projections for the next season.

Peterson improves upon these websites' projections by using a weighted average of fourteen such "black box" websites' projections in which he assigns weights based on each websites' historical accuracy. With this approach, he is able to achieve an R^2 of .569 and an $MASE$ of .479, an improvement of the individual website projections.

Others have attempted to assist fantasy football players in other ways via the application of more complex machine learning algorithms. Boris Chen from the New York Times applies a *Gaussian mixture model* to an aggregation of expert ranking data provided from FantasyPros.com to find clusters or tiers of players within the ranking data to help fantasy football players understand the natural tiers of NFL players [3].

Niltin Kapania from Standford applies both *linear regression* and *k-means clustering* to attempt to predict running backs' total season fantasy points and achieves results that are nearly on par with fantasy football expert Mike Kruger's running back projections [7].

Matt Bookman, a graduate student from Stanford attempts to predict weekly quarterback fantasy points by training both linear regression and support vector machine models to achieve a slightly greater Pearson's ρ than Yahoo's projections, indicating consistently better rankings each week than Yahoo's rankings [1].

Dr. J.J. McKinley uses the *random forests* machine learning algorithm, trained on each players' regular season fantasy point production to find the top value plays for each weekend's wild card games [9].

These varying models help to better serve fantasy football competitors in their league's performance in different ways; we wish to form our own projections for the quarterback position season total points to aid in preseason drafting and see if we can match or improve upon these results.

4 Data Description

The majority of our data was retrieved from [2] which includes NFL statistics broken down by each player for each week's game from the years 2008 to 2014. These data provides us with essential quarterback, wide receiver, and running back performance statistics from the 2008 - 2014 seasons. For the quarterback position, this includes the player name, the player's

respective team, the opposing team, their quarterback rating for the particular game, their fantasy-point production statistics, and other performance statistics. We then calculate each quarterback's *actual* fantasy points each week by using the *standard scoring* formula:

Fantasy Points = $\frac{1}{25}$ (Passing Yards) + 6 (Passing Touchdowns) + $\frac{1}{10}$ (Rushing Yards) + 6 (Rushing Touchdowns) -2 (Interceptions) -2 (Fumbles)

Next, we acquire teams' various seasonal defensive and offensive measures/projections from Football Outsiders. We merge the team offense statistics with each player's team and opposing team passing defense statistics with each player's opposing team to have a better understanding of how a player's own team's offense and opposing team's defense will influence their fantasy point production.

Because our main goal is to project season total fantasy points for each quarterback, we summarize each quarterback's performance statistics, games played, quarterback rating, *actual* season total fantasy points, and their team/opponents respective offense and defense statistics for each entire season.

We form the **training set** by labeling all quarterbacks' season fantasy points from years 2008 to 2012 as their *old points* to explain their subsequent season's performance from years 2009 to 2013, as their *new points*. Thus, we will train on these players' previous season total points to predict these players' upcoming season total points. Our **testing set** then consists of all quarterbacks' old fantasy points in 2013 which our models will use to predict the new fantasy points for 2014. Because we have the actual 2014 data, we can then ascertain our models' performance on the testing set by comparing our projected or fitted values to actual fantasy points the quarterbacks accumulated throughout the 2014 season.

5 Methodology

To choose the best players to draft to our teams we wish to predict the players' accumulated season fantasy points. To accomplish this, we choose to implement a *simple linear regression* model, what is often considered the "bread and butter" prediction method for statisticians and data scientists [12]. We first train a simple, one variable linear regression model that uses players' season totals of last year's fantasy points as the explanatory variable to predict their current year's season totals of fantasy points (the response variable).

After analyzing the results of this simple model, we can utilize more of the variables present in our data to hopefully increase the fit of our linear regression model.

To better understand which data might serve as explanatory variables, we can make exploratory plots to pick out which variables show correlations or trends with the dependent variable.

Next, we implement many *multiple regression* models which now include various combinations of explanatory variables such as:

- Quarterback rating

- Fantasy Point production statistics
- Opposing team passing defense statistics
- Player's team offense statistics

To quickly evaluate our models' performance, we will use the three following statistics:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$MASE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The best of our models would be the model which achieves the highest R^2 value while maintaining both low $MASE$ and $RMSE$ values on the **testing set**. Note that even "good" scores on the training set do not necessarily indicate an accurate forecasting model; a high performance on the training set coupled with a low performance on the testing set indicates that our model is *overfitting* the data [6].

To better visualize our models' performance, we plot the predicted points vs the actual points while using a loess smoother [4] to quickly characterize our scatter plots. We may also compare how close our results line up on the $y = x$ or the *line of perfect prediction* [12].

6 Implementation Details

After making heavy use of the `dplyr` package to summarize and filter our data by seasons, and form our training and testing sets. We use the built in R function `lm` to form our various linear models.

We train our first simple linear regression model in `R` and our two most successful multiple regression models via

- `modelOne <- lm(newFanPoints ~ oldFanPoints, data = TRAIN)`
- `modelTwo <- lm(newFanPoints ~ oldFanPoints + OppPassingDef, data = TRAIN)`
- `modelThree <- lm(newFanPoints ~ oldFanPoints + OppPassingDef + Status, data = TRAIN)`

In addition to checking our scatter plots for unusual residual patterns and visually measuring our goodness of fits, we call the `summary` function on our models to better analyze our p-values and coefficients of our linear models.

7 Analysis

To better understand our explanatory variables and their effects on the current season's fantasy points, we read the output from the `summary` call in `R` on our models. We include the output from our final most successful model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-172.23  -40.96  -12.41   35.10  318.33

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    10.51836    6.58286   1.598 0.110925
oldFanPoints                    0.68298    0.04143  16.486  < 2e-16 ***
OppPassingDef                   0.52096    0.07536   6.913 2.07e-11 ***
StatusMissing previous data   -33.55683    8.61913  -3.893 0.000117 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

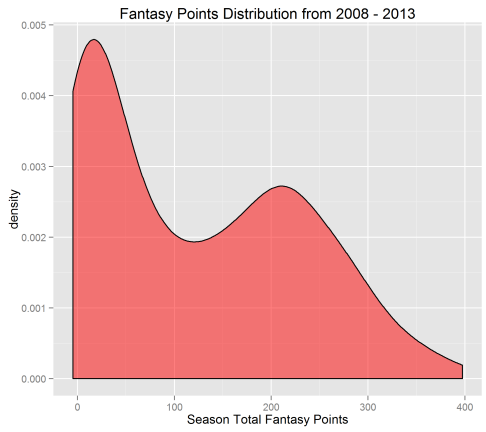


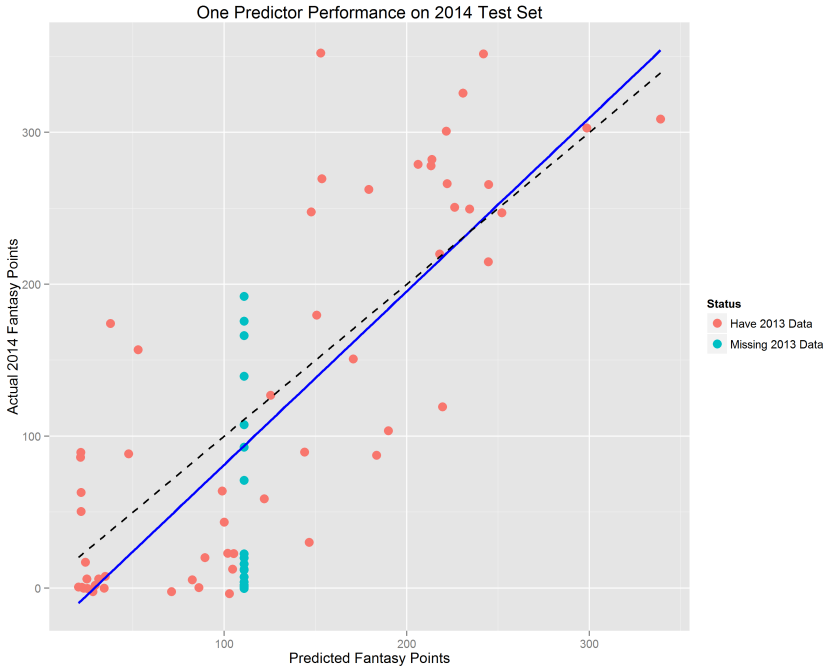
Figure 1: Bimodal Distribution

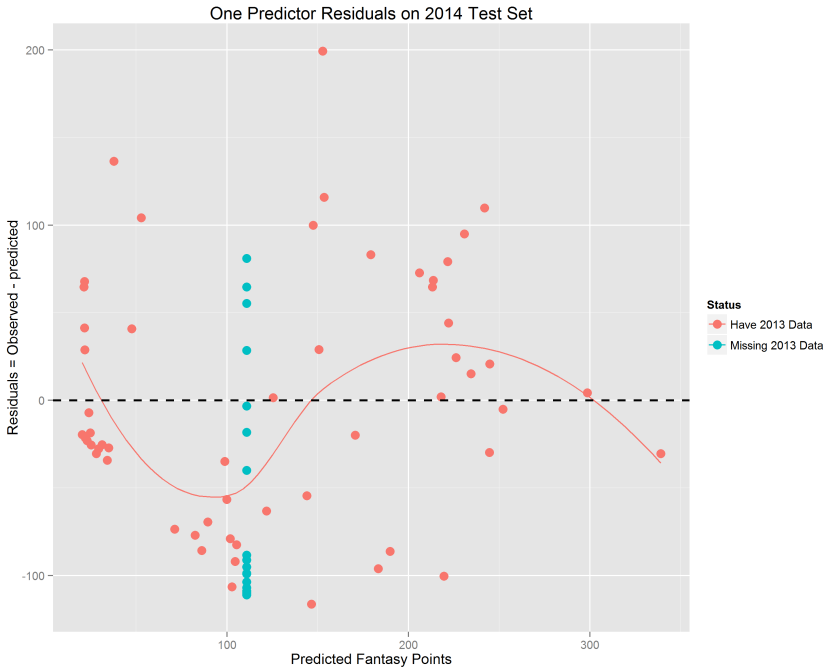
The most important part of regression analysis is analyzing our residuals [6]. We can see that while we might expect one or two high magnitude residuals as -172.23 and 318.33, more alarmingly we see a median value that is quite far from zero. This suggests that on average we are over predicting the players' fantasy points. A likely reason for this might be due to the *bimodal distribution* of fantasy points [11]. This suggests that we ought to model the weaker quarterbacks separately from the high-performing star quarterbacks. Unfortunately, none of our attempts to model these populations separately significantly improved our models' results.

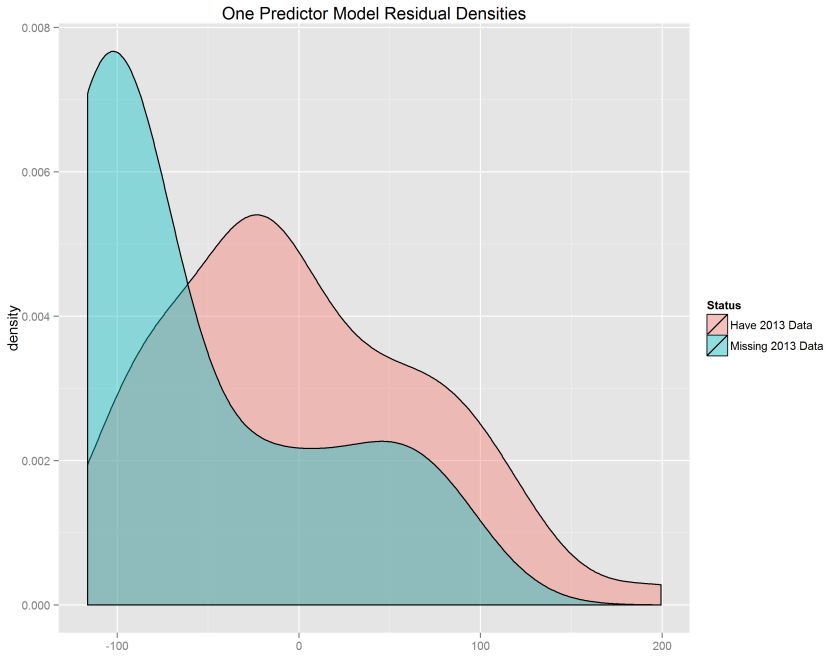
Note that the Standard Errors of `oldFanPoints` and `OppPassingDef` are at least an order of magnitude lower than the coefficient estimates, which indicates that our coefficient estimates are likely close to the true coefficient values. Also, the low p values corresponding to the explanatory variables are relevant to predicting the `newFanPoints` variable. Also of interest is that the estimated coefficient of the third variable indicates that quarterbacks whom do not have previous NFL experience are penalized by about 33 fantasy points for their upcoming year. This suggests new quarterbacks score less fantasy points than experienced ones. As expected, the largest contributor to predicting fantasy points comes from the `oldFanPoints` variable which should be interpreted as "for each 1 unit increase of last year's fantasy points, we predict an increase of about 0.68 for this year's."

8 Results

We include three plots for our simple and three-predictor regression models. The first depicts the goodness of fit, the second helps us see the residuals more closely, and the third depicts the distribution of the residuals. Note that we would ideally obtain normally distributed residuals centered at zero [8].

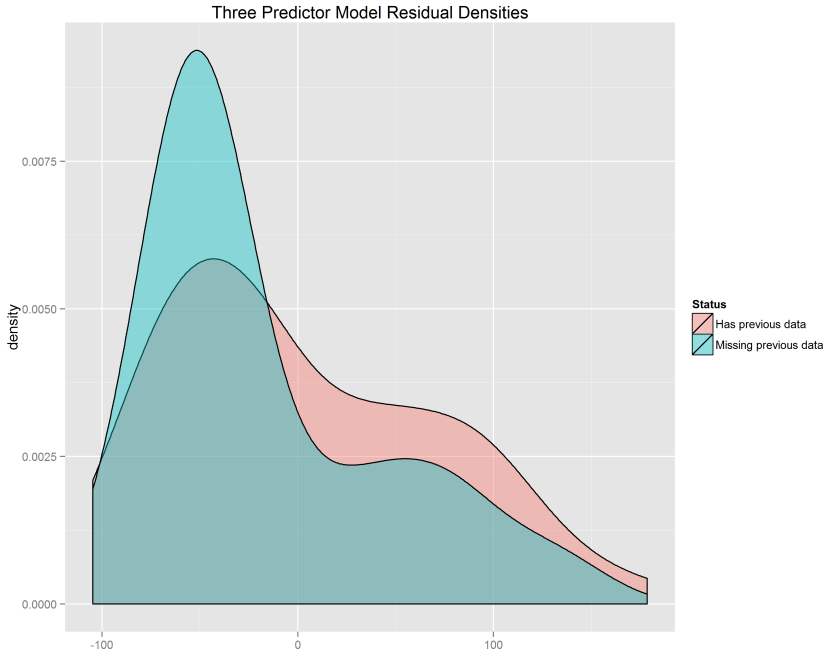






Compare those results with our best multiple regression model.


Three Predictor Residuals on 2014 Test Set
Residuals = Observed - predicted
100
0
-100
0
100
200
300
400
Predicted Fantasy Points
Status
Has previous data
Missing previous data


Three Predictor Model Residual Densities
density
0.0075
0.0050
0.0025
0.0000
-100
0
100
Status
Has previous data
Missing previous data

We summarize the results of our three linear models against Fantasy Football's projections in the following table.

Model	R^2	$MASE$	$RMSE$
Simple LM	.5497	.5068	75.352
2 Predictor LM	.6061	.4830	70.466
3 Predictor LM	.648	.455	66.575
FFA Projections	.569	.479	

Table 2: Model Evaluations: Includes players with *missing* 2013 data

9 Conclusion

The most surprising results from our modeling process was noting the power of *simple* linear regression. Using each players' previous season points (and filling in players' missing previous season points with the median of that year) got us to explaining 55% of the variation of the next year's points. Adding in the opponents' passing defense statistics improved our fit by an additional 6 percentage points. Finally, by training our model to deal with players that had missing data, we were able to explain up to 65% of the variation; this appears to be a significant improvement upon FFA's projections.

Unfortunately, data that would be available before the upcoming season such as a quarterback's age, quarterback rating, and projected team offense statistics turned out to lack any real predictive power. Considering that there is high amount of variability for our explanatory variables, it is surprising that we obtained the a significantly higher R^2 than FFA. However, as our still high $RMSE$ indicates, the standard deviation of unexplained variance is about 70.5; we might interpret this to mean that about 66% of the actual results will be within a range of $\pm$ 66.5 fantasy points of our predicted values. This means that there is still a lot of room for improvement.

Before moving on to other machine learning methods we could expect to easily improve our linear models by acquiring data for the players with missing season 2013 data (possibly through their college games). We would also probably benefit from training on more than just one previous season's fantasy points for the players that have several years of experience. Finally, we have yet to really solve the issue of the bimodal distribution of quarterback fantasy points. More research about dealing with multimodal distributions might help us in separately modeling the two populations of mediocre and star quarterbacks.

References

[1] Matt Bookman. *Predicting Fantasy Football - Truth in Data*. 2012. URL: http://cs229.stanford.edu/proj2012/Bookman-PredictingFantasyFootball.pdf (visited on 10/06/2015).

[2] John Broberg. *Excel for Fantasy Football*. 2014. URL: https://excelfantasyfootball.wordpress.com/get-free-nfl-stats-in-excel/ (visited on 09/28/2015).

[3] Boris Chen. *Turning Advanced Statistics Into Fantasy Football Analysis*. 2013. URL: http://www.nytimes.com/2013/10/11/sports/football/turning-advanced-statistics-into-fantasy-football-analysis.html?ref=football&_r=1 (visited on 10/06/2015).

[4] Robert Cohen. *An Introduction to PROC LOESS for Local Regression*. URL: http://www.ats.ucla.edu/stat/sas/library/loesssugi.pdf (visited on 10/06/2015).

[5] Koehler Hyndman. "Another look at measures of forecast accuracy". In: (). URL: http://www.robjhyndman.com/papers/mase.pdf (visited on 10/06/2015).

[6] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN: 1461471370, 9781461471370.

[7] Nitin Kapania. *Predicting Fantasy Football Performance with Machine Learning Techniques*. URL: http://cs229.stanford.edu/proj2012/Kapania-FantasyFootballAndMachineLearning.pdf. 2012.

[8] Mark Lunt. "Introduction to statistical modelling: linear regression". In: *Rheumatology* 54.7 (2011), pp. 1137–1140.

[9] J.J McKinley. *Using Machine Learning to Create Daily Fantasy Football Projections for the Wild Card Round*. 2014. URL: http://rotoviz.com/2014/12/using-machine-learning-create-daily-fantasy-football-projections-wild-card-round/ (visited on 10/06/2015).

[10] Isaac Peterson. *Fantasy Football Analytics*. URL: http://fantasyfootballanalytics.net/. 2015. (Visited on 09/28/2015).

[11] Erhard Reschenhofer. "The bimodality principle". In: *Journal of Statistics Education* 9.1 (2001).

[12] Nina Zumel, John Mount, and Jim Porzak. *Practical data science with R*. Manning, 2014.