Predicting Rookie Year Success of NFL Running Backs using College Career and Scouting Combine Metrics

Jacob Pate
(advised by Dr. Eliana Christou)

Abstract

Predicting the performance of potential players is one of the most important tasks a football team performs – it has a direct impact on the success of the team. Though each team's analytics models are closely held proprietary secrets not released to the public, each player's game statistics and combine performance is publicly available data. Our goal is to take the game and NFL combine statistics of *college running backs* and predict their first-year success in the NFL. We built a web scraper to assist in the bulk collection of statistics for running backs who played in college between 2011 and 2019. We chose to omit the 2020 and 2021 college seasons because the worldwide coronavirus pandemic caused disruptions to both college and professional football schedules, where less games were played as a result. We collected college and combine statistics for each player, as well as each player's first-year statistics in the NFL. Various models were investigated and suggested that a player's rushing attempts, receptions, receiving yards, and 40-yard dash time were important factors that determine players' success in NFL.

1. Introduction

We are interested in predicting the success of rookie-year NFL running backs using their college statistics along with their scores from the NFL scouting combine. We use several linear mean and quantile regression models for our predictions. We collect players' career college statistics and their metrics from the NFL Scouting Combine to use as predictors.

To better understand the context around our topic, we review several articles in relevant literature. First, we take Aaron Schatz's 2005 paper *Football's Hilbert Problems*, which detailed several issues with NFL game data collection and the way data at the time was being used (and not used). Schatz went on to manage one of the first NFL player stats websites and partnered with FOX Sports in 2005 and 2006 to perform analytics for the broadcasts.

We also read Jack Porter's 2018 paper *Predictive Analytics for Fantasy Football: Predicting Player Performance Across the NFL*. Their model assumes no correlation between a player's different metrics, which made us want to look at how those metrics may be correlated. Because they focus solely on *fantasy* sports, their goal is boil down performance to a single number that predicts fantasy points.

Jordan Cook's *The Relationship Between the NFL Scouting Combine and Game Performance Over a Five Year Period* was especially illuminating. Until reading this paper, we hadn't considered using a player's NFL Combine statistics in the prediction model. This paper suggests

that a running back's performance in the 40-yard dash and their professional performance are strongly correlated, and we also found that metric to be a significant predictor in almost every model.

Finally, we review 2005's *Quantile Regression* by Roger Koenker to learn more about quantile regression. Specifically, quantile regression is used as an alternative to mean regression, especially when the error term is heteroscedastic, and it allows you to obtain a more complete picture of the conditional distribution as one can focus on different quantiles. Moreover, quantile regression is more robust to outliers and for that reason, can be more appropriate when we deal with some extreme performances of players.

2. Models under Consideration

For the analysis we consider two types of models: mean regression and quantile regression models. *Mean regression* models the relationship between the conditional mean of the response and the covariates, while *quantile regression* models the relationship between the conditional τth quantile of the regression and the covariates, where $\tau \in (0, 1)$. For this work, we use $\tau = 0.5$, which corresponds to *median regression*. Recall that, for a univariate response variable Y and a p-dimensional vector of predictors $\boldsymbol{X}$, a linear mean regression model is given by

$$E(Y|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p.$$

Similarly, a linear median regression model is given by

$$Q_{0.5}(Y|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

where $Q_{0.5}(Y|\boldsymbol{X})$ denotes the 0.5 conditional quantile of the response given the covariates.

We also consider a nonparametric quantile regression model, which does not assume a linear relationship between parameters but instead allows the predictor to form an equation based on the data. We use a *local linear quantile regression* technique to locally fit lines across the data.

Since the number of predictors under consideration is large (see Section 3 for details), we consider various *variable selection and dimension reduction techniques*. Specifically, we use stepwise variable selection, principal component analysis (PCA), and sliced inverse regression (SIR; Li 1991). A brief discussion of these methods is given below.

2.1. Stepwise Regression Model

Stepwise Regression is a hybrid variable selection algorithm between Step-Up (Forward) Selection and Step-Down (Backward) Selection. In stepwise selection, the algorithm starts with the predictor that has the highest R^2 value. Then, at each 'step' a new predictor is added to the model based on how significantly that predictor increases the R^2 value of the predictors in the model. Every time a new predictor is added to the model, the algorithm rechecks *every* predictor in the model to ensure each predictor is above a given threshold of significance. If a predictor is then determined to be nonsignificant, it is dropped from the model.

For most of our stepwise models, we keep six to seven significant variables. However, our predictions for year 2016 are worth noting here – in both quantile and linear mean stepwise

regression models, we keep nine significant variables. All kept variables for each year are further explored in Section 3.

2.2. Principal Component Analysis

PCA is an *unsupervised* dimension reduction technique, which means the data it takes in are unlabeled, i.e., the response variable is not taken into consideration. PCA was first introduced by Pearson (1901), although the name seems to originate in the influential paper by Hotelling (1993). The idea is to replace the $p \times 1$ predictor vector $\boldsymbol{X}$ with a $d \times 1$ predictor vector $\boldsymbol{B}^T\boldsymbol{X}$, where $\boldsymbol{B}$ is a $p \times d$ matrix of weights, $d \leq p$, such that $\boldsymbol{B}^T\boldsymbol{X}$ maintains the maximum variability of $\boldsymbol{X}$. Specifically, we are looking for a $p \times d$ matrix $\boldsymbol{B}$ that solves the following problem:

$$argmax_{\boldsymbol{C}}\ var(\boldsymbol{C}^T\boldsymbol{X}).$$

The column vectors of the matrix $\boldsymbol{B}$ are called the principal components (PCs) of $\boldsymbol{X}$, which are independent of one another and are ranked in order of importance. Generally, the number of principal components is chosen based on how many components explain at least 70% of the variability.

For this analysis, we use $\boldsymbol{B}^T\boldsymbol{X}$ as the new sufficient predictors to fit linear mean and quantile regression models. It was observed that two or three principal components were necessary for the different years considered; see Section 4 for more details.

2.3. Sliced Inverse Regression

Sliced Inverse Regression (SIR), introduced by Li (1991), is a *supervised* dimension reduction technique that utilizes a *weighted* PCA using an inverse regression curve. The idea is to regress each coordinate of $\boldsymbol{X}$ against Y. This will give a $p \times d$ matrix $\boldsymbol{B}$ that defines the new reduced sufficient predictors $\boldsymbol{B}^T\boldsymbol{X}$. As before, we use these new predictors to fit linear mean and quantile regression models.

For our models, we choose a suitable number of the dimension d based on the "elbow method" defined in the plot of the number of dimensions (or components) used in the model, as demonstrated below in Figure 1. More discussion on our results is given in Section 4.

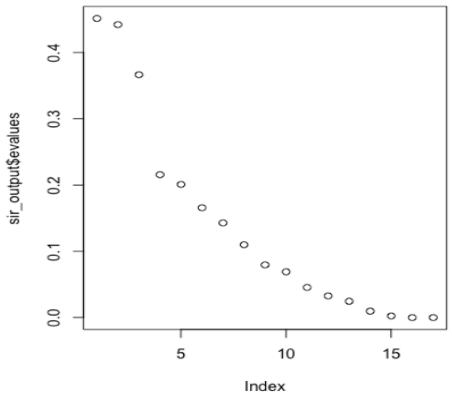

Figure 1. Eigenplot to determine the number of sufficient predictors.

3. Data

Our data consist of observations from 2011 to 2019. The 17 predictor variables consist of players' overall college career statistics, NFL combine statistics, and each player's team running percentage (i.e., how often a team executes a running play vs a passing play). A detailed list of the predictor variables is given below:

- Games Played
- Rushing Attempts
- Rushing Yards
- Yards per Carry
- Touchdowns
- Receptions
- Receiving Yards
- Yards per Reception
- Team Run Percentage
- Height
- Weight
- 40-Yard Dash
- Vertical Jump
- Bench Press
- Broad Jump
- Cone Exercise
- Shuttle Run

The data was collected from several sources. Specifically, the players' college metrics were obtained from the official NCAA website, their corresponding professional metrics were obtained from the official NFL website, and the run percentage of each of the represented teams was obtained from teamrankings.com. Both the official NCAA website and the official NFL website lack an export feature to aid in bulk collection of statistics, so we built a web scraper to grab the information for us.

As far as the response variable, we investigated different metrics to quantify NFL performance of a player, but we found that *Yards per Carry* (YPC) or *Yards per Reception* (YPR) give the best results. For that reason, we decided to run each of our models twice; one with YPC as the response variable and one with YPR as the response variable. Finally, to deal with heteroscedasticity of the response variable, we decided to use the log-transformation.

We start by fitting the full models for mean and quantile regression. Although the detailed analysis is presented in Section 4, we provide a brief discussion on which predictor variables were significant.

In our full models, the following variables were significant throughout the years:

- Yards per Reception
- 40-Yard Dash
- Height (in the linear model only)

To deal with multicollinearity, we continue by fitting the stepwise linear mean regression, where the following variables were selected throughout the years:

- Attempts
- Receptions
- Receiving Yards
- Height
- 40-Yard Dash

As far as the stepwise linear quantile regression model, the significant variables throughout the years were:

- Receptions
- Receiving Yards
- Yards per Reception
- 40-Yard Dash

To measure the performance of each model, we use the mean absolute error, given by

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

where y_i denotes the true response variable, $\hat{y}_i$ denotes the estimated response, and n is the number of players for which we make predictions in any given year.

4. Results

For our analysis, we predict NFL performance for years 2016-2019. Specifically, we use historical data to fit the model and then obtain the specific year's predicted response. For example, to predict NFL performance for year 2016, we use the data from 2011 to 2015 to fit the model. Then, we use the predictor values for year 2016 to predict the response variable for that year. We repeated this procedure for all years and all methods.

The methods under consideration are:

- Linear mean regression with all predictor variables (Full Model)
- Linear quantile regression with all predictor variables (Full Model)
- Stepwise linear mean regression
- Stepwise quantile mean regression
- Mean regression with predictors resulted from PCA (PCA linear model)
- Quantile regression with predictors resulted from PCA (PCA quantile model)
- Mean regression with predictors resulted from SIR (SIR linear model)
- Quantile regression with predictors resulted from SIR (SIR quantile model)
- Local linear quantile regression model

Table 1 demonstrates the significant variables for each model and both response variables under consideration. Note that it does not make sense to include significant variables for the PCA or SIR models, as the components that make up those models are linear combinations of *every* predictor in the system.

We bring attention to a few interesting things – when we make predictions on a player's YPR, the most significant predictors are their vertical and horizontal jumping capabilities. In both quantile models we see touchdowns and games played are significant as well. When we predict for a player's YPC, their height and 40-yard dash times are most significant. This is likely intuitive. What we find of particular interest is that often, a player's *ground* statistics will be significant when predicting their *catching* metrics, and vice versa. We see in every model with YPC as a response that YPR shows up under significant variables. Across both stepwise models with YPR

as a response, rushing yards, receptions, and YPC are significant, which we believe is not as intuitive.

	2019	2018	2017	2016	all models
Stepwise linear model (YPR)	Games Played Rushing Yards YPC Receptions Height Vertical Jump Broad Jump	Rushing Yards YPC Receptions Height Vertical Jump Broad Jump	Games Played Rushing Yards YPC Receiving Yards Weight Vertical Jump Broad Jump	Rushing Attempts Rushing Yards Receiving Yards YPR Weight 40-Yard Dash	Rushing Yards YPC Vertical Jump Broad Jump
Stepwise quantile model (YPR)	Games Played Rushing Yards Touchdowns Receptions Weight Vertical Jump Broad Jump	Games Played Rushing Yards Touchdowns Receptions Height Vertical Jump Broad Jump Cone Exercise	Games Played Rushing Yards Touchdowns Receptions Weight Vertical Jump Broad Jump Cone Exercise	Receptions Receiving Yards Run Percentage Weight Bench Press	Games Played Rushing Yards Touchdowns Receptions Weight Vertical Jump Broad Jump
Quantile model (YPR)	Games Played Touchdowns Vertical Jump Broad Jump	Games Played Touchdowns Broad Jump	Games Played Touchdowns Broad Jump	Rushing Yards Weight	Games Played Touchdowns Broad Jump
Linear model (YPR)	Games Played Vertical Jump Broad Jump	Vertical Jump Broad Jump	Games Played Weight Vertical Jump	Weight 40-Yard Dash	Vertical Jump

	2019	2018	2017	2016	all models
Stepwise linear model (YPC)	Rushing Yards Receptions Receiving Yards YPR Height 40-Yard Dash	Attempts Receptions Receiving Yards YPR Height 40-Yard Dash	Games Played Attempts Receptions Receiving Yards YPR Height 40-Yard Dash	Games Played Attempts Receiving Yards Height 40-Yard Dash	Receptions Receiving Yards YPR Height 40-Yard Dash
Stepwise quantile model (YPC)	Touchdowns Receptions Receiving Yards YPR 40-Yard Dash	Rushing Yards Receiving Yards YPR Height 40-Yard Dash	Rushing Attempts YPR Height 40-Yard Dash Vertical Jump	Rushing Attempts Receptions Height 40-Yard Dash	Receptions YPR Height 40-Yard Dash
Quantile model (YPC)	Receptions Receiving Yards YPR 40-Yard Dash	Receiving Yards YPR	YPR 40-Yard Dash	40-Yard Dash	YPR 40-Yard Dash
Linear model (YPC)	Receptions Receiving Yards YPR Height 40-Yard Dash	Receptions Receiving Yards YPR Height 40-Yard Dash	Games Played Receptions Receiving Yards YPR Height 40-Yard Dash	Games Played Height 40-Yard Dash	Receptions Receiving Yards YPR Height 40-Yard Dash

Table 1. Significant predictor variables for each method for the different years.

Table 2 presents the mean absolute error for all different models and for both response variables under consideration. We observe that models with YPC as a response variable had the lowest error of all, but models with YPR as a response are also sufficiently low in error. Moreover, we find for models with YPC as a response variable that the stepwise linear model performs best over the four years for which we make predictions, and for the models with YPR as a response variable the PCA quantile model performs best over the same period.

Model Type (log) YPC	Error per Player 2019	Error per Player 2018	Error per Player 2017	Error per Player 2016	Sum of Average Absolute Error
stepwise linear model	0.137554068	0.111630136	0.186712855	0.140242796	0.576139856
stepwise quantile model	0.141501611	0.132357278	0.189010907	0.112539214	0.57540901
PCA linear model	0.144070856	0.136457614	0.186210506	0.137991156	0.604730131
PCA quantile model	0.150932131	0.122782754	0.185891314	0.135771158	0.595377357
sliced inverse linear model	0.14730171	0.151459592	0.19854691	0.166103843	0.663412055
sliced inverse quantile model	0.140932854	0.127363328	0.189568235	0.163399703	0.621264122
nonparametic quantile model	0.138768135	0.132542533	0.201100225	0.166263145	0.638674038
quantile model	0.139801286	0.129115081	0.199060209	0.119453171	0.587429746
linear model	0.142977537	0.127556449	0.194996856	0.132186526	0.597717368

Model Type (log) YPR	Error per Player 2019	Error per Player 2018	Error per Player 2017	Error per Player 2016	Sum of Average Absolute Error
stepwise linear model	0.342880564	0.194830384	0.423034836	0.353850435	1.314596219
stepwise quantile model	0.327688851	0.19424739	0.392143097	0.298118606	1.212197944
PCA linear model	0.303955823	0.19213488	0.406945844	0.384141662	1.287178209
PCA quantile model	0.29254583	0.173244295	0.386777456	0.287014154	1.139581735
sliced inverse linear model	0.322529403	0.196447668	0.439861755	0.344277009	1.303115834
sliced inverse quantile model	0.299064197	0.180324973	0.39560652	0.278052045	1.153047736
nonparametic quantile model	0.296803042	0.167987056	0.379528931	0.321065877	1.165384906
quantile model	0.332763681	0.203101399	0.370727742	0.335601055	1.242193876
linear model	0.332763681	0.203101399	0.370727742	0.335601055	1.242193876

Table 2. Mean absolute error for the different models and both response variables.

5. Discussion

It is interesting to see how the significant variables for each response are made up both of variables one may predict from the start, such as jump distances for YPR response and sprint speed for YPC response, but also from things we might not guess, such as rushing yards or yards per carry for YPR response and receptions and receiving yards for YPC response. They are exactly the opposite from what we may have assumed before starting this project.

What may be of interest in future work is tracking throughout the games in a season or throughout the seasons of a player's career how well their cumulative metrics in these measurements predict future performance.

Although our research does not look specifically at fantasy sports, it would be interesting to see how running backs picked based on their projected YPR and YPC perform in a fantasy league.

We did not consider for this paper a combination of the two responses, but combining them may result in a richer picture of success for an NFL running back, who would likely need to perform at a high level both in running the ball on the ground and running it after catching it in order to be successful long term in the league.

Of note also is that our models predict performance basically in vacuum, so results may not translate directly to true gameplay. Running backs are one position in a game with eleven members on offense and eleven members on defense. It is obvious that all 22 athletes on the field during a given play influence the outcome of that play, and all throughout the game. Another interesting area could be looking at each position in relation to the other positions in play at any given time, though this would likely result in a very complex model.

References

Cook, Jordan, 2019. The Relationship Between The NFL Scouting Combine And Game Performance Over A Five Year Period. *Georgia Southern University, Electronic Theses and Dissertations*. 1906.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417-441, and 498-520.

Koenker, R, 2005. *Quantile Regression*. Cambridge University Press.

Li, K.-C., 1991. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316-327.

Pearson, K., 1901. On Lines and Planes of Closest Fit to System of Points in Space. *Philosophical Magazine*, 2(11):559-572.

Porter, Jack W., 2018. Predictive Analytics for Fantasy Football: Predicting Player Performance Across the NFL. *University of New Hampshire, Honors Theses and Capstones.* 406.

Schatz, Aaron, 2005. Football's Hilbert Problems *Journal of Quantitative Analysis in Sports: Vol. 1: Iss. 1, Article 2.*