PENALIZED LOGISTIC REGRESSION TO ASSESS NFL QUARTERBACK PERFORMANCE

4/26/2016

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document. When you're ready to add your content, just click here and start typing.]

Peters, Benjamin G

[Email address]

Introduction

The premier professional American football league is the National Football League (NFL). Over the past decade, there appears to have been a large emphasis in the quarterback position. For example, NFL quarterbacks are often assigned a win-loss record similar to that of baseball pitchers or hockey goaltenders. The quarterback position is the only position in the NFL to be assigned a win-loss record. As can be seen in Figure 1, the emphasis on the quarterback position is understandable.

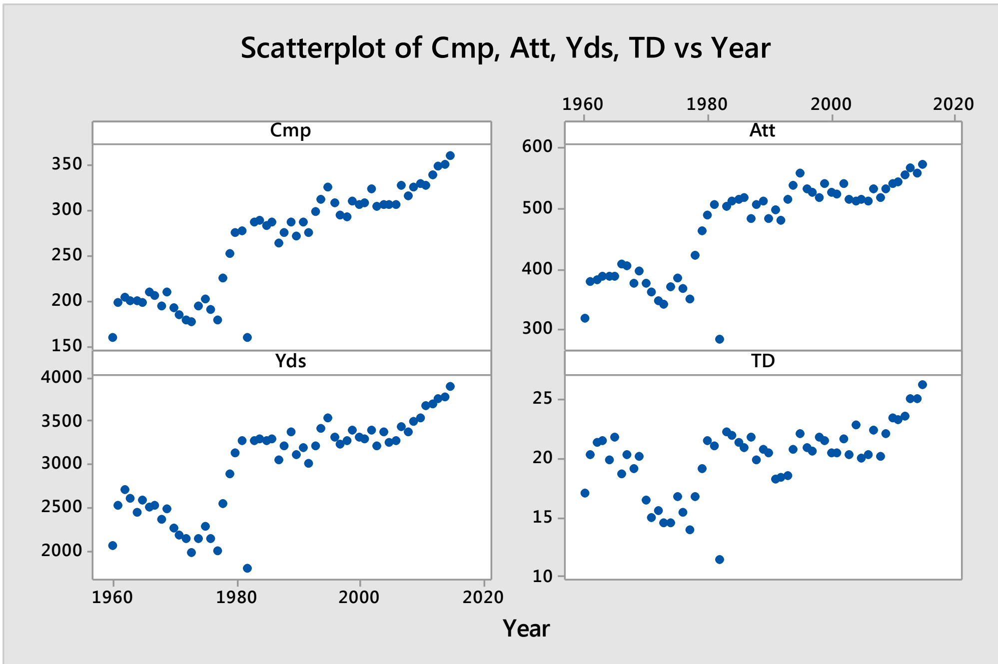

Figure 1: Team Yearly Average

Figure 1 shows that over the years there has been a general trend in the yearly passing attempts, completions, yards and touchdowns per team (It should be noted that the 1970s have a reputation for being a time period when the running backs were most prominent. As a result, teams passed much less in this decade than in any other decade since 1960).

With the increasing trend in passing, it makes sense that quarterback would be the highest paid position in the league. In the NFL, quarterbacks average over $3.8 million. With that much money being paid to quarterbacks, there is heavy responsibility on a team's quarterback to perform. The goal of this study is to find quarterback performance metrics (pass completion percentage, yards, touchdowns, interceptions, etc.) that are indicators of a team's probability of winning a game and use these important metrics to determine a quarterback's contribution to their team's chances of winning.

There have been a few metrics introduced to measure quarterback performance. This paper discusses two of them. The first is the passer rating [1]. The passer rating was adopted by the NFL in 1973. It is a function of pass completions per attempt, passing yards per attempt, passing touchdowns per attempt, and interceptions per attempt. The metric ranges from 0 to 158.3. The metric fails to account for game situations and for other variables the quarterback may be responsible. The other metric is Total Quarterback Rating (Total QBR) [2]. This metric was designed by ESPN's Stats & Information Group and is designed to measure how a quarterback's play contributes to scoring points and winning. ESPN has not released an actual formula for this metric. While the metric has the benefit of easy interpretability (scale 0-100), its computation is very detailed and is dependent on data not necessarily obtainable to the general public. In this study, a penalized logistic regression is used to determine weights for several common game statistics. These weights can then be used on a team's per-game averages to determine a value which reflects the team's chances of winning a game. This value is broken up between the quarterback and the rest of the team based on how responsible the quarterback is for certain statistics. As a result, the quarterback's contribution to their team's chances of winning can be determined. The method has the benefit of being easy to use while utilizing all the statistics a quarterback is directly responsible for.

The next section provides a brief overview of logistic regression and penalized logistic regression. The model is then developed. Then using the fitted coefficients of the regression model, the aforementioned value is developed and a small demonstration with 4 teams is shown. The report ends with some ideas of improvement.

Logistic Regression

In ordinary least squares regression, given a set of continuous observations $y_i \in R, i = 1, \dots, n$ and a feature matrix $X \in R^{p+1}$, the objective is to fit a model to the conditional expectation of the response given the set of features. Typically, the model is assumed to be of the linear form $\beta_0 + \sum_{j=1}^{p} \beta_j x_j$. In order to fit this model, the least squares approach looks to minimize the residual sum of squares which results in the closed form solution $\boldsymbol{\beta} = (X'X)^{-1}X'\boldsymbol{y}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ and $\boldsymbol{y} = (y_1, \dots, y_n)'$.

In logistic regression, the response is now binary $(0,1)$. Therefore, the normal least squares approach is not appropriate as it assumes the response can take on any real value. Instead of modeling the response, we can model the probability that the response takes on one of either two options. However, probabilities are limited to being between 0 and 1. Therefore least squares is still not appropriate. Instead we can use the logit transform of the conditional probability that the response is 1 given a set of features. If we let $p(x) = P(Y = 1|X = x)$, we can model the logit of $p(x)$ as:

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

Solving for $p(x)$ yields

$$p(x) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)\right]}$$

Typically, the parameters $\boldsymbol{\beta}$ are fitted by maximizing the log-likelihood function:

$$l(\boldsymbol{\beta}) = -\sum_{i=1}^{n}\left[(1 - y_i)\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right) + \log\left(1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)\right)\right]$$

While no closed form solution exists, the parameters can be fit using Newton's method or gradient descent among other iterative approaches. However, the interest is in performing variable selection. Therefore, a slight modification is made.

Penalized Logistic Regression: Elastic Net

Just as in ordinary least squares regression we can apply a penalty to the log-likelihood function such as:

$$l^*(\boldsymbol{\beta}) = -\sum_{i=1}^{n}\left[(1-y_i)\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j\right) + \log\left(1 + \exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j\right)\right)\right] - \lambda P_\alpha(\boldsymbol{\beta})$$

where $P_\alpha(\beta) = \frac{1-\alpha}{2}\left|\left|\boldsymbol{\beta}\right|\right|_2^2 + \alpha\left|\left|\boldsymbol{\beta}\right|\right|_1$ [3].

It is not difficult to see that when $\alpha = 1$, the penalty is the same as the LASSO penalty. As $\alpha \to 0$, the penalty acts like the ridge parameter. The result of maximizing this penalized log-likelihood equation is β_j's that shrink to zero as λ grows. The α term acts to shrink parameters that are correlated. In the problem for this project, there are several features and some are correlated. Therefore, this penalty shall reduce the number of significant features while shrinking the coefficients of the remaining features to account for correlations.

Data Collection

For this project, 170 games were sampled over the course of 5 seasons (2010-2014). The NFL season is broken down into 17 weeks. Every team plays 16 games and they have one bye week (a week off). For each of the 17 weeks, a winning team and a losing team were chosen randomly. Data was collected from Pro-Football-Reference.com [4] and The Football Database [5]. Initially, 46 features were considered. These features include pass yards, rushing yards, passing touchdowns, etc. The full list can be seen in Table 1 on the following page. The response y_i is the result of the game for the team of interest. Due to the method of sampling, there are an equal amount of wins (1) and losses (0). It should be noted that the chance of a tie is possible. However, there were no ties in the sampled games. If there were, the ties would be lumped with the losses and be considered 'Not Win'. After creating a 170 row matrix with 46 feature columns, the features are standardized into z-scores so that the variable selection removes features based on their importance and not because of relative scale. For example, the number of passing touchdowns will be much smaller than passing yards since it is typical for a team to pass for more than 200 yards while scoring 2 touchdowns. Therefore, the coefficient for yards may be very small as the unit change in a yard would not lead to a large change in the probability of winning compared to the unit change of a touchdown.

Regression Model

In order to fit the penalized logistic regression, there needs to be a choice of the penalty terms α and λ. α is chosen to be 0.75 in order to lean towards variable selection. The LASSO parameter is chosen using 10-fold cross validation. The MATLAB function 'lassoglm' performs this operation. MATLAB looks to find the λ which minimizes the deviance which is the same as maximizing the log-likelihood. The result is shown in the Figure 2.

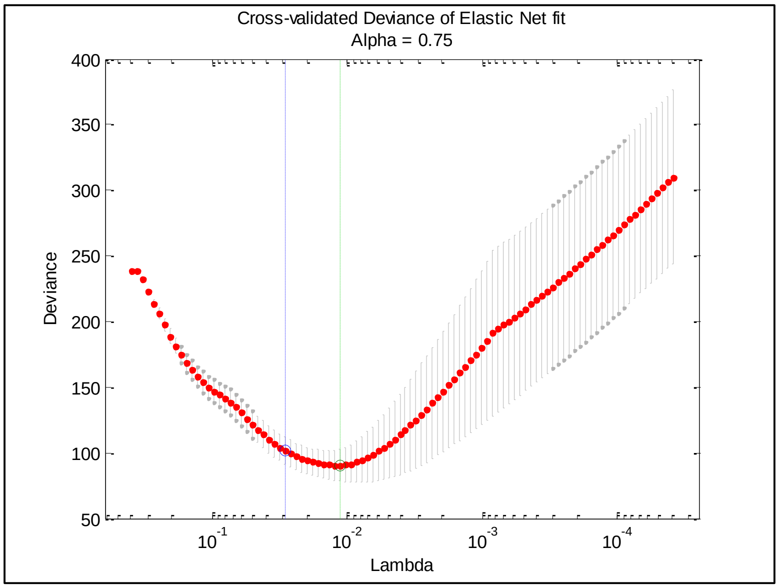

Figure 2: Cross Validation

The result is $\lambda = 0.0114$. The model is refit using the two parameters. The results are shown in the Table 1:

Table 1: Features and their Coefficients

Feature	β	Feature	β
Pass Comp	0.000	Avg. Net Yardage	0.000
Pass Att	-0.328	Punt Downed <10	0.000
Pass Yards	0.549	Opp Punt TD Ret	-0.119
Pass TD	0.921	Punts Bk	-0.192
Sacks Taken	-0.229	# of Punts Forced	0.000
Sk Yards Lost	-0.517	Avg. Return Ydg	0.159
Rush Att	0.617	Punt Ret TD	0.000
Rush Yds	0.000	Bk Punts Forced	0.000
Rush TD	1.035	Touchback %	0.000
Turnovers	-0.391	Kick Yds/Return	0.000
Opp TO Return TD	-0.320	Opp Kick Yds/Ret	0.000
Opp Pass Comp	0.000	Kickoff TD	0.195
Opp Pass Att	0.217	Opp Kickoff TD	0.000
Opp Pass Yds	0.000	FG Made	0.489
Opp Pass TD	-1.399	FG Att	0.000
Sacks Forced	0.000	Opp FG Made	-0.927
Sack Ydg	0.000	Opp FG Att	0.000
Opp Rush Att	-0.325	Blocked FG	0.182
Opp Rush Yds	-0.147	Opp Bk FG	0.000
Opp Rush TD	-0.859	Safety	0.283
Turnovers Forced	0.320	Opp Safety	-0.187
Turnovers Ret TD	0.164	Pen Yds	-0.221
# of Punts	-0.018	Opp Pen Yds	0.226

If the exponential of the coefficients is taken, the result can be interpreted as the change in the odds of winning for a unit change in the variable. Since the data is standardized, $\beta_0 = 0$. With the coefficients fitted, the next step is to use these coefficients to measure the influence a quarterback has on the result of a game.

Quarterback Influence

From Table 1, the significant features related to the Quarterback position are *Pass Att*, *Pass Yards, Pass TD, Rush Att, Rush TD, Turnovers, Opp TO Return TD*, *Opp Safeties*, and *Penalties*. Note that the sacks features could be attributed to the quarterback but for this study, they are attributed to the offensive line (5 players protecting the quarterback). It is not surprising that scoring touchdowns (*TD*) would lead to a higher chance of winning. However, it is interesting that Passing Attempts (*Pass Att)* is negatively weighted. This could be due to the fact that when teams fall behind they tend to pass more frequently in an attempt to catch up. As a result they have a large amount of passing attempts, but they still lose. While rushing attempts, yards, and touchdowns are typically associated with the running back position, quarterbacks are also quite involved in this part of the game as well. While few will contribute a large amount, the rushing production of the quarterback should be considered. The reasoning behind the strong weight for the rushing attempts relates to the tendency of teams relying on the running the ball while they are ahead to keep the clock ticking down.

In order to separate the influence of the quarterback on the outcome of the game from that of the rest of the team, a proportion of each feature is assigned to the quarterback based on their average production relative to the team's average production. Take this excerpt from the data as an example:

	Attempts	Yards	Pass TD	Sacks Taken	Sk Yards Lost	Rush Att	Rush TD	TO
NE	609	4921	33	21	134	438	13	13
Tom Brady	582	4291	34	26	170	36	0	12
Percentage	0.956	0.872	1.030	1.238	1.269	0.082	0.000	0.923

The top row shows the names of the features. The second row shows the 2014 season totals for each feature and the third row shows quarterback Tom Brady's 2014 season totals. The final row is the proportion of each feature assigned to Tom Brady. In the analysis, the totals are divided by 16 to yield the per-game average. These per-game averages are standardized using the means and standard deviations for each respective feature from the 2010-2014 game data. Then the appropriate proportion of each standardized feature is allocated to the quarterback. Using the weights (β_j's), a linear combination, $X\boldsymbol{\beta}$, with the standardized features for each team, and by extension each quarterback, can be calculated. For simplicity, this is defined as '*Score*'. A '*Score*' greater than 0 is favorable as it corresponds to the chance of winning being greater than 0.5. A negative value indicates either sub-par passing statistics or a propensity to commit turnovers or worse having those turnovers returned for touchdowns. By subtracting the total team '*Score*' by the quarterback's '*Score*', the remaining team contribution is determined. Since the quarterback does not play defense nor do they kick field goals, the remaining team contribution consists primarily of the defense and kicking features along with the rest of the rushing features. A graphical representation of this is shown in Figure 3.

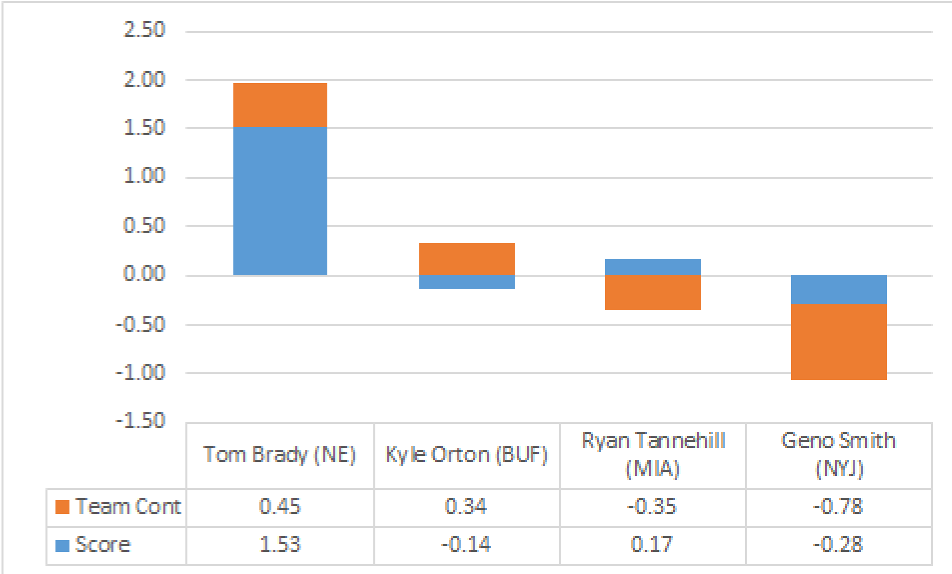

Figure 3: AFC East Scores

The graph shows the results for the 4 teams from the Eastern division of the American Football Conference of the National Football League. Colloquially, this is known as the AFC East. It consists of 4 teams: the New England Patriots, Buffalo Bills, Miami Dolphins, and New York Jets. The results indicate how quarterback play can affect a team's win probability. The obvious example is Tom Brady's large score contributing to his team's large overall score. On the other hand, Kyle Orton's performance appears to hinder his team who otherwise has a rather good score. The score and remaining team contribution can be added together and converted to a probability.

$$p(x) = \frac{1}{1 + \exp[-(X\boldsymbol{\beta})]}$$

For these four teams the results are shown in Table 2:

Table 2: Summary of AFC East

	Team Score	P(W)	Exp Wins	Act Wins	Error
New England	1.976	0.878	14.051	12	-2.051
Buffalo	0.195	0.549	8.776	9	0.224
Miami	-0.172	0.457	7.315	8	0.685
N.Y. Jets	-1.056	0.258	4.128	4	-0.128

The '*Score*' can provide a way for teams to assess their quarterback. It can also be used by the players themselves to gauge their relative contribution. For players performing at a high level, it is expected that they would have a high '*Score*'. Thus the metric can be used as a bargaining tool when the time comes for contract negotiations.

Conclusion

As time has progressed, the focus on the passing game has increased immensely. Whereas running backs were seen as the premier position in the 1970s, that title now belongs to the quarterback. On average, they are the highest paid players in the league and the statistics have been steadily increasing. The focus of this project was to 1) determine the metrics recorded during an NFL game that are most important for predicting whether a team wins and 2) use these important metrics to determine a quarterback's contribution to their team's chances of winning. This metric can be used in several ways. It can be used to rank the quarterbacks providing a list for those that are interested in discussions of which player is better. More importantly, it can be used by teams and players alike to measure performance which plays a part in the contracts the players sign.

There are some drawbacks to the overall design of the model. The model does not account for game situation. A player can accumulate yards or even touchdowns late in the game when the result is decided and that will count the same as if they had done when the result of the game had not been decided. In addition, the data collected for the graph in the previous section does not account for games missed. For example, Geno Smith of the New York Jets did not play in two games during the 2014 season. Therefore, his averages are only for fourteen games not sixteen. In order to account for this, the overall team values should only include the games that he played in. However, it is more challenging to get the data filtered in such a way.

In order to expand on this project, a more detailed look at player statistics can be used. Factors ignored such as game situation can be used to weight performance. Using the example of the player accumulating stats that do not affect the game, those stats can be weighted lightly to give a fair representation of the player's accomplishments. In addition, the finances of the player can be accounted for. Based on performance, the amount of money a player is expected to be paid when it is time for a new contract can be studied using the '*Score*' defined in this study.

References

[1] "Passer Rating." *Wikipedia*. Wikimedia Foundation. Web. 26 Apr. 2016.

[2] "Total Quarterback Rating." *Wikipedia*. Wikimedia Foundation. Web. 26 Apr. 2016.

[3] "Documentation." *Lasso or Elastic Net Regularization for Generalized Linear Model Regression*. Web. 26 Apr. 2016.

[4] "Pro Football Reference." *Pro-Football-Reference.com*. Web. 26 Apr. 2016.

[5] "Football Statistics and History | The Football Database." *FootballDB.com*. Web. 26 Apr. 2016.