SS 3859 Final Report

by

NICHOLAS SERRATORE

KAREN AMY STEIL

DARIO MELCONIAN

**Table of Contents**

1. Acronyms and Abbreviations…………….………..…………….…….………………….3

2. Introduction……………………………………………………………………...………...4

3. Methods…………………………...………………………………………….....................7

4. Results………………………………………...………………..……………………...…12

5. Conclusion.…………...……………………………………………….…………………14

6. Discussion………………………………………………………………………...……...16

7. Works Cited………………………………………...………………...……...…….…….18

**Acronyms and Abbreviations**

The list provided below describes the abbreviations used throughout the report. All the terms related to the ‘Time In Pocket’ dataset from PFF (Pro Football Focus) are defined in detail below. It should be noted that ‘More’ and ‘Less’ added to the predictors is a result of being in the pocket less than or more than 2.5 seconds, as these statistics were broken into 2 categories.

player Football Player Name

POS Position of the Player

G Number of Games Played

DB Dropbacks – the number of times the QB dropped back to pass

SK Sacks – the number of times the passer was sacked (tackled by defender while in possession of the ball for a loss of yards)

COM Completions – the number of times the passer completed a pass

TD Touchdowns – the number of touchdowns by the passer

NFL Passer’s NFL Passer Rating

INPK Average Time in Pocket

X2RUN Average Time to Run/Scramble (move around before throwing ball)

DB\_perc Percentage of Dropbacks with Ball Held for Less/More than 2.5 seconds

X2ATT Attempts – the number of times the passer threw the ball

COM\_perc Completion Percentage – the percentage of completed passes to pass attempts

INT Interceptions – the number of interceptions thrown by the passer (passes thrown and caught by the opposing team)

**Introduction**

Throughout the turn of the 21st century, the National Football League (NFL) has increasingly become more involved with statistics and data analysis. In sports in general, the aspect of creating the best probability to win any given game has always been the objective. In recent history, however, with technology developing at an exponential rate, sports industry franchises have hired data analysts and statisticians to develop models to comprehend all the data observed in a game, further innovating the idea of achieving main goal in sports - to win.

Statistical analysis for NFL data at the ‘quarterback’ position has been increasingly significant for all sports data analysts. NFL data scientists have tremendous roles in the NFL to the point that every franchise has a team of statistical/data analysts creating models in the pursuit to extract conclusions to ultimately increase the probability of winning any given game. This is easily seen in the data from this [research paper](https://rc.library.uta.edu/uta-ir/bitstream/handle/10106/26766/KING-THESIS-2017.pdf?sequence=1). Regression analysis was performed on quarterback data; in particular, height, age, completions, sacks, NFL passer rating, etc. Numerous predictors used in the analysis of this students’ data are similarly used in the model in this report. This shows that the relevance of these statistics for quarterbacks are relevant and widely used. Conclusions are drawn to further assist team staff and other personnel, as well as analysts and other professionals in an attempt to predict future outcomes. The analysis of sports data, specifically, NFL data, has developed into its own industry and has grown exponentially in recent years. Thus, the analysis of this data is extremely relevant in today’s data-driven world.

In the game of football, there are 11 players aside competing against each other to reach each other’s endzones to score ‘touchdowns’ (TD’s). The result of a touchdown is 6 points. There are 2 ways to achieve this; running with the football and throwing it to receivers downfield. The field of play is comprised of 100 yards of grass, with an endzone on each end. To commence any given play, a player must snap (or hut) the ball through his legs from the ground to the passer (his teammate). Although there are multiple positions in the sport, the focus on this project is on the passer. The passer in football is labeled as the ‘quarterback’ (QB for short). He sits in the ‘pocket’ (area where the ball tends to be thrown from, and where defensive players aim to tackle the QB from) as he makes his decision to run the ball himself, pass it to a receiver downfield, or hand the ball off to a runner. In regard to this project, the focus is on the passer’s time in the pocket. This includes any time between the snap and the throw.

The actions of the QB throughout the play as well as the decisions he makes are what help generate the statistics that will be analyzed in this report.

These statistics include:

* his completions
* completion percentage
* attempts
* dropbacks
* dropback percentage
* sacks
* average time to sack
* touchdowns
* NFL passer rating
* Interceptions
* average time to run/scramble
* average time in pocket
* average time to throw
* average time to sack.

Some of these factors are separated by being in the pocket for less than/greater than 2.5 seconds. This totals the number of predictors to 22. These factors are known to be significant in any NFL analysis on the quarterback position. Thus, the goal of this analysis is to use regression to determine the significance and influence of each of these 22 factors on the number of touchdowns (TD’s) scored by any given QB.

All data in this project was taken from PFF (Pro Football Focus). This website focuses on thorough analysis of the NFL. Data was taken from this credible source from the 2015 season to the 2019 season. Over 500 QB’s seasons were included in this dataset. The focus in this report is to analyze what statistics influence or do not influence the number of touchdowns scored by a quarterback. The conclusion of this data can be beneficial to coaches and analysts as well as staff members of any NFL franchise. In analyzing this data, conclusions help gear focus towards specific aspects of the game of football, to essentially increase the rate of touchdowns for a given quarterback, to ultimately increase the probability of winning more games.

**Methods**

To effectively analyze and interpret this data, multiple linear regression was performed

using the PFF quarterback data. It should be noted that in NFL football, there are occurrences where a football player throws a small number of passes throughout the entire NFL season, sometimes as low as a single pass. The data from these players is ineffective in contributing to the analysis of the influence of quarterbacks on touchdowns. For instance, a player who is normally a receiver but had one game played as a QB that threw one pass in an entire season, has inflated statistics in the categories considered in this project. Moreover, if a backup QB replaced a starter for one game, and carried similar inflated statistics, the same skewing would apply. Thus, it was decided that for the best analysis and interpretation of the data, all players with 2 or fewer games played were eliminated from the dataset, and not included in the analysis of this project. This approach to cut out known outliers or inflated skew factors helps effectively perform regression on the data at hand. Prior to developing a multiple regression model, a correlation analysis was performed to determine which variables are significantly correlated with the dependent (response) variable, TD:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Correlation with TD | P-value | Variable | Correlation with TD | P-value |
| DB | 0.8883 | 0.0000 | Less\_SK | 0.3636 | 0.0000 |
| G | 0.8665 | 0.0000 | More\_ATT | 0.8669 | 0.0000 |
| INPK | -0.1247 | 0.0546 | More\_COM\_perc | 0.4033 | 0.0000 |
| Less\_ATT | 0.8495 | 0.0000 | More\_DB\_perc | -0.0967 | 0.1369 |
| Less\_COM | 0.8630 | 0.0000 | More\_DP | 0.8567 | 0.0000 |
| Less\_COM\_perc | 0.3722 | 0.0000 | More\_INT | 0.5544 | 0.0000 |
| Less\_DB\_perc | 0.0966 | 0.1372 | More\_NFL | 0.5629 | 0.0000 |
| Less\_DP | 0.8492 | 0.0000 | More\_SK | 0.6761 | 0.0000 |
| Less\_INT | 0.4432 | 0.0000 | X2ATT | 0.0016 | 0.9803 |
| Less\_NFL | 0.6398 | 0.0000 | X2RUN | 0.0336 | 0.6057 |

The table above shows that variables DB, G, Less\_ATT, Less\_COM, Less\_COM\_perc, Less\_DP, Less\_INT, Less\_NFL, Less\_SK, More\_ATT, More\_COM\_perc, More\_DP, More\_INT, More\_NFL, and More\_SK are significantly correlated with the dependent variable since their p-values are less than the significance level of 0.05. Therefore, those variables were included in the multiple regression model:

Here is the summary of the model:

Table

Description automatically generated

The summary of the model shows that there are multiple predictors that are insignificant, and therefore a stepwise forward selection method was applied to select the best variables. This method starts with an empty model, tests the addition of each variable based on the p-value and adds the variable with the smallest p-value. This process continues until none of the new variables improve the model to a statistically significant extent. Here are the results of the process:

Table

Description automatically generated![Chart, scatter chart

Description automatically generated]()

This shows that all the variables selected by the stepwise forward selection method are significant at the 5% significance level (α = 0.05). However, the residual plot shows that there is a clear quadratic pattern in the residuals, which violates the assumption of linearity. Upon inspecting the scatterplots of individual predictors against the dependent variable, it was found that there is a quadratic pattern in the variables DB and Less\_NFL:

![Chart, scatter chart

Description automatically generated]()![Chart, scatter chart

Description automatically generated]()

Therefore, quadratic terms for those two variables were added to the model and the residual plot was again generated:

Table

Description automatically generated ![Chart, scatter chart

Description automatically generated]()

The residual plot shows an improvement, however there are still some outliers present with standardized residuals greater than two. Additionally, the Cook’s distance was calculated to check for influential observations and 20 records were found to be influential. Upon removing the outliers and influential observations, we get the following results:

Table

Description automatically generated ![Chart, scatter chart

Description automatically generated]()

**­­­Results**

The residual plots show an approximately random pattern, which means that the linearity assumption is satisfied. The homoscedasticity assumption is also satisfied since the variance of residuals is approximately constant. To check the normality assumption, we generate a histogram of residuals and a normal Q-Q plot:

![Chart, histogram

Description automatically generated]() ![Chart, line chart

Description automatically generated]()

The histogram and the Q-Q plot show that residuals are approximately normally distributed, however there is some skewness in the tails of the distribution. To assess these analyses, a Shapiro-Wilk test, and a Breusch-Pagan test were conducted to further test these assumptions. Using the Shapiro-Wilk test, the p-value was found to be 0.4141, which confirms that the normality assumption is not violated. Using a Breusch-Pagan test, the p-value was found to be approximately 0.0002, which shows that the equal variance assumption is not violated. The model has an of 0.946, which implies that 94.6% of variation in the number of touchdowns is explained by the independent variables in the model.

Chart, scatter chart

Description automatically generated

Additionally, it was important to include visualizations of the significance of the model assumptions. Seen here, in the reduced preferred model, the model assumptions are not violated. The linearity assumption is not violated as the values are equally distributed above and below the line. In other words, the values in the residual plot are approximately equidistant from the 0 line. Moreover, the model assumption for equal variance was met. This is seen as the values from fitted value of 0 to 15 do not concentrate at a specific residual value and are evenly spread throughout the graph. Moving across the fitted values from 0 to 15, the height of the values does not very across. This means that the equal variance holds, as previously numerically proven through the BP test above.

**Conclusion**

The data shows that the variables: DB (dropbacks), DB2, Less\_NFL (NFL passer rating under 2.5 seconds in the pocket), Less\_NFL2, More\_NFL (NFL passer rating over 2.5 seconds in the pocket), More\_SK (sacks after more than 2.5 seconds in the pocket), More\_COM\_perc (completion % when in pocket more than 2.5 seconds), Less\_COM\_perc, and More\_INT (interceptions after being in pocket more than 2.5 seconds) are the most significantly correlated predictors with the dependent variable. These predictors are the factors that have the strongest influence the number of TD’s that a quarterback scores over the course of a season. The factors with positive correlations are: DB, DB2, Less\_NFL2, More\_NFL, and, More\_INT. As these factors increase, so does the number of TDs. Contrarily, the factors with negative correlation to the response variable are: Less\_NFL, More\_SK, More\_COM\_perc, and Less\_COM\_perc. As these factors increase, the number of TD’s decreases.

It was interesting to conclude that some factors such as ATT (attempts), INPK (average time in pocket), and 2SK (average time to sack) were not significant in the influence of TD’s scored by a given quarterback. It should be noted that these factors are all influenced by many other players (offensive players’ protection, defence players’ pressure). It is hypothesized that perhaps these factors have more influence on an entire team’s offensive production, such as yardage or total points, as well as an entire team’s defensive production throughout a game, through yards against or points against. Thus, these predictors don’t exactly influence the production of TD’s by a quarterback.

The conclusions drawn from this report help further pinpoint the most crucial factors in scoring touchdowns. It can be concluded that the positively correlated predictors consisting of dropbacks, NFL passer rating over and under 2.5 seconds in the pocket, and interceptions over 2.5 seconds in the pocket are essential statistical factors in the number of touchdowns scored. Quarterbacks with a higher passer rating tend to throw more accurate passes, resulting in a greater number of touchdowns. Additionally, the greater number of dropbacks by a given QB, the greater number of chances for deep passes to a receiver downfield. Lastly, in regard to interception’s after being in the pocket more than 2.5 seconds and its positive correlation with touchdowns scored, this seems to merely be a reflection of downfield pass attempts. It can be concluded that the longer pass attempts by a QB, the greater the chances of an interception, as well as the greater chances of a touchdown being scored, compared to a short 5-yard pass that is rarely intercepted and rarely resulted in a touchdown. Thus, the greater the number of dropbacks, the greater number of deep ball passes, which increases the chances of interceptions in any given play. Thus, a QB who throws a high volume of interceptions also completes a significant amount of deep ball passes, which in turn, results in TD’s scored. This is the explanation as to why interceptions having been in the pocket for longer than 2.5 seconds positively influences the number of TD’s scored. It is simply broken down into deep ball attempts having sat in the pocket greater than 2.5 seconds, to ultimately increase chances of TD’s, which is caused by dropbacks, resulting in a high interception rate. Thus, it must be considered that increasing interceptions also increases points scored against, as an interception results in the change of possession.

**Discussion**

Through the in-depth analysis of the data, coaches can target specific areas of improvement for their quarterbacks. For example, a coach can conclude that the total number of passing attempts is not related to TD’s scored. This can help coaches focus on what to improve on when working with their quarterbacks. Additionally, it helps a coach focus on other aspects of an offence, such as influential factors like NFL passer rating, interceptions, and dropbacks. Or, contrarily, focusing on minimizing factors such as sacks (More\_SK), and putting less emphasis on the importance of completion %’s in the pocket (COM\_perc), can further assist the production of touchdowns. Coaches can further develop those aspects of their quarterback’s game in an attempt to raise the probability of TD’s scored.

Additionally, this data also shows that interceptions positively correlate TD’s, which is something coaching staff must carefully consider. This factor is one that coaches do not want to increase, as it also increases the points against, which negatively affects the probability of winning a game. It makes sense because the more dropback time, the more time to allow receivers to run downfield. This concept increases the probabilities of interceptions, as the football is thrown in desperation and has more time to be tracked by the defence. Additionally, however, this concept also increases the probabilities of touchdowns scored due to the receiver being further downfield and closer to the endzone, as well as having more space and being open. Thus, coaches must strategically consider increasing dropback time, while carefully considering the ratio between interceptions resulting in points against versus touchdown’s resulting in points for. These two factors are a result of a high number of deep pass attempts, which comes from a high number of dropbacks in the pocket for more than 2.5 seconds. This is seen through the combined positively correlated predictors. Thus, NFL coaching staff members must minimize sacks when in the pocket for over 2.5 seconds and worry less on the importance of completion percentage in their quarterbacks. In light of this, they must also focus on increasing the NFL passer rating, as well as increasing the number of dropbacks, while carefully limiting the “number of interceptions resulting in points against versus the number of touchdowns scored” ratio, with their quarterbacks. All of this must be considered and thoroughly focused on to help achieve the goal in scoring touchdowns, ultimately increasing the probability of winning a football game.

**Works Cited**

The dataset used to perform the multiple linear regression analysis in this project can be found here:

* [dataset](https://premium.pff.com/nfl/positions/passing/time_in_pocket?position=QB&season=2019&week=1%2C2%2C3%2C4%2C5%2C6%2C7%2C8%2C9%2C10%2C11%2C12%2C13%2C14%2C15%2C16%2C17)

Additionally, the statistical analysis performed on NFL players in the quarterback position by Nicholas King can be found here:

* [research paper](https://rc.library.uta.edu/uta-ir/bitstream/handle/10106/26766/KING-THESIS-2017.pdf?sequence=1)