A Statistical Analysis of the Simpson College Football Offense

Erik Hall

August 25, 2016

Abstract

This project explores the predictability of play-calling of the Simpson College football offense during the years of 2012-2014. The models used to assess the predictability include simple analysis as well as linear and logistic regression.

1 Introduction

In the years of 2012 and 2013, the Simpson College football team achieved records of 6-4 and 7-3 overall. Both of these seasons marked the highest win totals since the start of head coach Jim Glogowski's career at Simpson. Then in 2014, with very high expectations from both the team itself as well as the Iowa Conference, the team posted a 3-7 record overall. The decline in success is easily attributed to a large amount of injuries as well as the graduation of a 4-year letter winner, all-conference, school record-breaking quarterback. While this is the the case, perhaps the decline in success can be due to other factors. This is where the statistical analysis may assist in the search for factors contributing to the decline in success. With the focus being on the offense, the goal is to determine if the play-calling and statistics associated with the play-calling were predictable. In other words the goal is to see if the play-calling tendencies of the seasons of 2012 and 2013 were predictable in a way such that the predictability had a negative effect for the season of 2014.

2 The Collection of Data

Many athletic teams found in the levels of high school, collegiate, and professional often take advantage of the use of filming their own practices and games in order to evaluate their players' ability. With the recent advancement of technology, these athletic programs now have access to software programs capable of storing the film recorded at games and practices. These software programs are also capable of storing relevant data associated with the given plays found on the film. The data has a large number of categories that plays can be organized into, and teams can pick and choose which categories to fill out when filming. In this application of sports statistics, the Simpson College football team uses a film software program known as Hudl. The data from Hudl used for this project

was entered by hand on the sidelines of football games. It's important to recognize that the manual entering of data will have some errors associated with it, and that this given case is no different. With that being said, due to errors some plays were not recorded and therefore were left out of the data collection.

3 A Further Explanation of Hudl

To give more context to what Hudl actually is, we can delve into it a little more. An example of something that can be seen on Hudl can be seen in Figure 1. Most obviously seen in the example of Hudl, is the video portion of the screen, but for the purposes of our analyses we are more concerned with what is directly below the video. As we look closer, we see a wide range of categories that each film clip is split into. These categories range from the the play number of the current game, to down and distance, as well as the play type. Note that in the yard line category has negative numbers as possible values. The explanation for this is that in terms of a teams possession, the yard line relative to which half of the field the team is on. In other words if a team is on its own half of the field, the yard line will be listed as a negative. Whereas if the team is on the opponents side of the field, the yard line will be listed as positive. For our given analysis of data, the categories describing plays found within Hudl will be used as the factors for prediction.

4 Methods for Prediction

In terms of measuring predictability of play-calling, there will be three main methods of analysis. The first being a simple statistical analysis of box score data, then we will look at quantitative analysis through linear regression, and lastly we'll look at logistic regression involving a qualitative analysis. Due to the many observations of variables, the computer software program R will be used for the calculations. The use of both linear regression and logistic regression require the data being analyzed to be split into a training set and a test set. The training set of the data is what each model is "trained" on. Whereas the test set of the data is what the model is "tested" on. Generally we are more concerned with the results of the test set, because the data being used is new to the model. Note that since we are concerned with the predictability of the 2014 season, the sets of data will be split up such that the 2012 and 2013 seasons will be used as the training set for models, and the 2014 season will be used as the test set.

5 Simplistic Statistical Analysis

Before we look more into advanced statistical analysis, it's important to recognize that sometimes simpler techniques can give more insight into the analysis of data. So with that being said, we can see a summary of data obtained from the box scores of games in the 2012, 2013, and 2014 seasons in Table 1. Note that the statistics found in the table are considered the scoring statistics, or in other words the statistics associated with the productivity of the offense. Given the differences in win totals, it is somewhat surprising to see similarity in the data

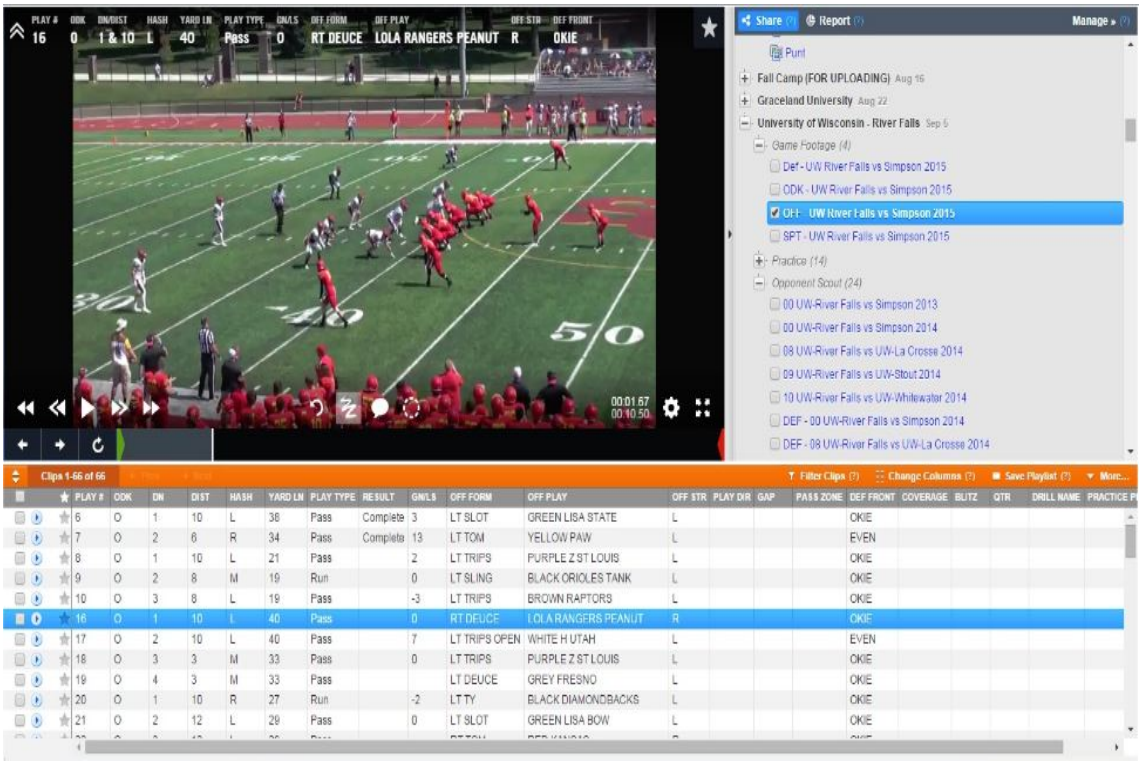


Figure 1: This is an example of what one could see when working with the film software program Hudl.

	2012	2013	2014
Average Starting Position (Yard Line)	33.19	31.81	30.74
Points Per Game	27.5	32.7	20.9
Average Yards Gained Per Play	5.06	5.16	5.30
Number of Turnovers	24	15	20
Red Zone Percentage	59%	77%	74%
3rd Down Percentage	46%	74%	34%
Average Time of Possession	30:49	32:22	31:20

Table 1: This is the summary of the statistical analysis done on the scoring statistics.

for each season. Even with the similarities, there are some glaring differences that separate the 2014 season from the others in a negative way. These differences are that the 2014 team only averages 20.9 points per game and that the team had a third down conversion rate of 34%. From a coach's perspective it's reasonable to expect a low third down conversion rate to lead to lower amounts of scores considering that the offense will find itself not being able to sustain drives and therefore not being able to score.

Given that we have this information there are a few things that we can take away giving us more insight for when we do more advanced analysis techniques. The most important item of information that can be taken away from the simple analysis is that the third down conversion percentage hints at predictable play calling. This is because a low third down conversion rate suggests difficult distances to go on third down, and an offense would have difficult third down situations if they are not successful the previous downs. The lack of success on previous downs can come from a defense watching film and learning tendencies on given downs, thus if a defense knows what a potential play could be, the defense can put the offense in situations that create low chances of success. Ultimately what we can take away from our simple analysis is that play-calling may be predictable, and thus is something to look into.

Before we look more into the advanced techniques, there is a simple logical explanation of changes in success that we can can consider. For the explanation we will look at the 2014 season only. The 2014 season began with a three game win streak against non-conference opponents, which was followed by a seven game losing streak against conference opponents. Using logic based off of film analysis done by a team, it can be somewhat expected that a team will do well in the majority of its non-conference games considering that the opponents are not played annually compared to conference opponents that are played annually. The reason that it can be reasonable to expect this possibility is that non-conference teams have little film of the their opponent available to them. This can be especially noticeable in the case of a team playing a non-conference team for the first time. This explanation is dependent upon the amount of film shared between coaches, but often times teams usually only have access to film of their own games and then games of the opponent played prior to their game. Ultimately the possible explanation is that Simpson had an ad-

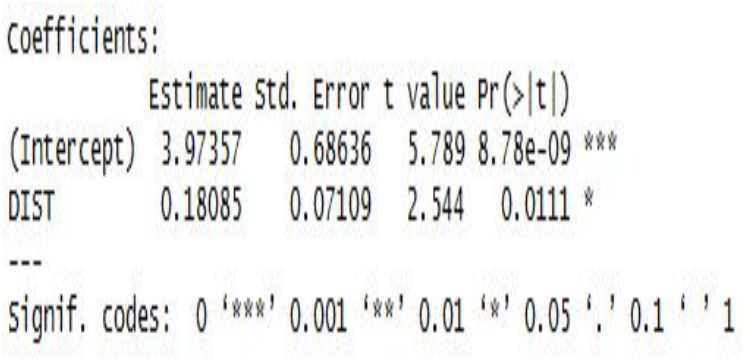


Figure 2: This is the output from R for the linear regression with distance as the predictor

vantage in non-conference games because, the opposing teams did not have the same opportunities as conference teams to analyze the predictability, in terms of film, of Simpson. Therefore Simpson may have been more likely to struggle in conference games.

6 Linear Regression

6.1 Distance Predicting Gain or Loss

With the limited number of quantitative variables found in Hudl, there have only been a few meaningful calculations. Of the most notable calculations that could be used in terms of predictability, is the linear regression predicting the gain or loss of a play with the predictor of distance to go on a given down. The results of the linear regression from R can be seen in Figure 2. Note that we see that distance has a significant p-value, suggesting the significance of the variable. The p-value significance codes can be seen at the bottom of the figure. For a given p-value to be at least somewhat significant, it's value must be at most 0.1. As the p-value becomes smaller, the variable associated with the p-value becomes more significant. For this linear regression, R gives the equation

$$\hat{y} \approx 3.97357 + 0.18085x.$$

Note this equation is for the seasons of 2012 and 2013. To give context to this equation, let us look an example. Consider a situation of 2nd down with 6 yards to go. In this case it is expected that the offense will gain about 5.05 yards, since we simply substitute the 6 in for x. Thus in this particular situation, the offense is not expected to gain the yards necessary to make a first down. Now with that being said, the given equation above expects the offense to gain a first down if $0 \leq x \leq 4$. Note that there are some limitations with this equation with an assumption being that the distance to go is discrete this is due to the data being discrete.

In terms of predictability, there's an important piece of information that can be drawn from this particular model. For the 2012 and 2013 seasons, if a defense could put the offense in a position in which more than 4 yards were needed for the first down, it could be expected that offense would not gain the needed yards. This suggests play-calling that doesn't have plays designed to gain more than 5 yards at a time.

While this is a very notable result, there are couple of issues that arise. First the predictability is related to our training set of the seasons of 2012 and 2013. This is an issue because we are more concerned with the predictability of the 2014 season, although it is important to note that earlier seasons can often times affect outcomes of later seasons in terms of play calling. The other issue that occurs is that when running a linear regression on the test set, distance is no longer a significant factor according to the significance codes of R. The output of the linear regression with distance predicting the yards gained or lost on the test set, can be seen in Figure 3. So while yes it is true that distance is no longer technically a significant predictor, there is something of note to consider.

If we further examine the output of the linear regression, we see that p-value for distance is close to being considered significant. With such a small difference preventing distance from being a significant predictor, it is safe to consider the predictor significant with a minor but important assumption. Recall that the predictor was significant for the training set, but also recall that the training set had more observations than the test set. The assumption is that if there existed more observations for the 2014 season, then distance would have a p-value that would be considered significant by R. We can go about verifying our assumption of significance by running a linear regression on the combination of the training and test sets. The R code output for the linear regression done the combination of the sets can be seen in Figure 4. As we review the output from R, we see that we now have a significant p-value of 0.00395. Something to note here is that the p-value we obtained is more significant than when we had run a linear regression on only the training set. This suggests that distance is a significant factor in the test set as well since the combination of the training and test set showed significance.

Now with all of this being said, it somewhat difficult to focus on what we were attempting to show in the analysis. In short what we've shown is that in each of the seasons we're looking at the yards gained by the offense on a given down could be predicted by the distance to go on a given down. Thus an opposing defense could be more inclined to bring pressures or blitzes on early downs knowing that offense will struggle with distances of 5 yards or greater to go.

7 Logistic Regression

7.1 Down Predicting Play Type

There are more meaningful calculations that can be done off of Hudl when using the qualitative variables. With that being said we can use logistic regression

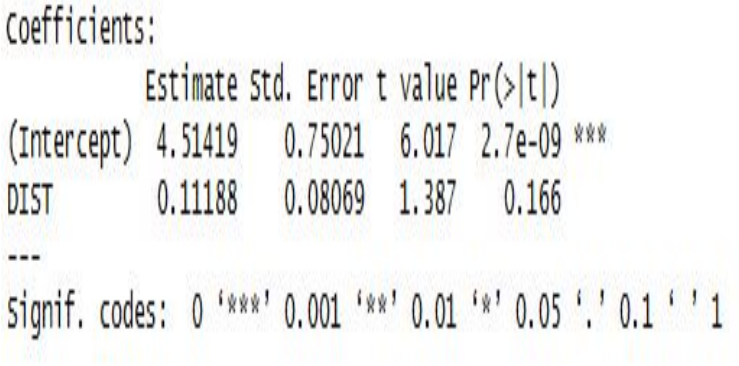


Figure 3: This is the output from R for the linear regression with distance as the predictor on the test.

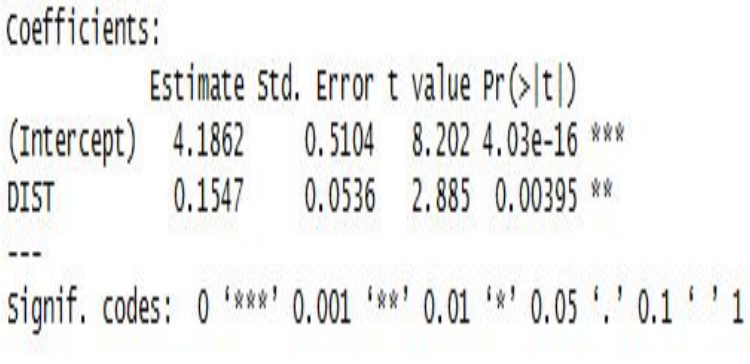


Figure 4: This is the output from R for the linear regression with distance as the predictor on the training and test set.

to look for other potentially predictable tendencies. The first that we will look into involves the use of down as a factor predicting play type. Before we run the regression, it is also worth note that a coach's intuition often says that down can predict run or pass. So we can go in with the expectation of seeing a significance of the down variable. similarly to when we ran linear regression earlier, we will first run the logistic on the training set. The output that R gives for the regression can be seen in Figure 5. Just like we saw in linear regression a low p-value indicates a significant variable, and in this case we see that down is a very significant variable for predicting play type. Note that the output from R is now a little bit different since we are dealing with qualitative variables. The difference is that the "Estimate" column of the output is different than when we had been dealing with linear regression. The reason for this difference is that the logistic regression model uses a fitting method called maximum likelihood whereas linear regression uses the least squares method. For the use of logistic regression, the maximum likelihood fitting method uses a logistic function to determine probability. The function that is used is,

$$P(x) \approx \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}.$$

For our given application of logistic regression, we simply substitute the values found in our output in for the function with x being the given down. Note that more than one predictor can be used in logistic regression, hence the addition of variables up to variable p. When we substitute the values for a first down situation our given probability is about 0.52. To give more context to what this value means, the value of 0.52 means that on first down it is predicted that the play type will be run. Then to find the probability of a pass play we simply take the complement of the value for a run, thus in this case the probability of a pass play on first down is about 0.48. The summary for the probability of run or pass on each down can be seen in Table 2. As we see the progression of downs it appears that the likelihood of a pass play occurring rises. From a coach's intuition this is somewhat expected because third down can generally be classified as a "passing down" since it is more likely to gain yards through passing rather than running. While this information is only good for the training set, it is important to note that the predictability of previous years can lead to predictability for current years.

Now that we have a good background for how logistic regression works well with the training set in terms of down predicting play type, we can shift our attention towards the test set. The output from R for the logistic regression of down predicting play type on the test set can be seen in Figure 6. We see again that down is considered a significant predictor for the regression, but an issue arises when calculations are attempted with the logistic function. The issues can most likely be attributed to the differences in the coefficients for the "Estimate" column. We see that in Figure 6 the intercept value is much greater than the intercept in Figure 5. These differences cause estimates of probabilities for run and pass plays to be skewed such that each down is predicted to have a probability of at least 0.95 for run plays. In terms of concluding whether or not down is considered to have significance is difficult. If we use the p-value of down though, we can consider the predictor to be significant with a warning that estimations of probabilities will be skewed. This can be something to look

Down	Probability of Run	Probability of Pass
First	0.52	0.48
Second	0.39	0.61
Third	0.28	0.72
Fourth	0.19	0.81

Table 2: These are the predicted probabilities of run or pass given the down.

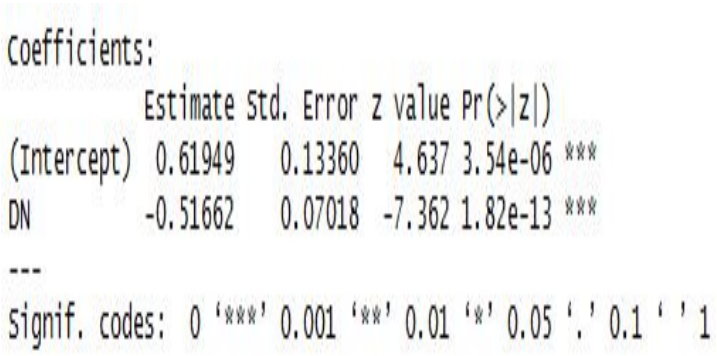


Figure 5: This is the output from R for the logistic regression with down predicting play type on the training set.

further into for future work.

7.2 Down and Distance Predicting Play Type

We have some idea that down may potentially predict the play type, even with possible inaccurate results for the test set. With that being the case we can make the regression a little more specific by adding distance as a predictor, so now we can look at specific situations. Just as we have done with the other regressions we will run the model on the training set first. The output for the logistic regression on the training set can be seen in Figure 7. Just as we saw when running the logistic regression with down as the only predictor we see that both down and distance are significant given their p-values.

Rather than showing a table of possible probabilities, we will look at a specific situation due to the amount of possible situations that exist. Consider the situation of third down and fifteen yards to go. Before we look at the results, we can come up with an expectation to see if our results match our expectation. Since there are a large amount of yards to be gained with a limited number of downs, we can expect a passing play more often than run in the situation. Substituting both 3 and 15 int our logistic function we get a probability of run as about 0.175, thus the probability of pass is 0.825. We see that our expectation of a pass play is met with the probability of a pass play at 82.5%. Similar to our previous logistic regression, we see significance for our predictors on the training set.

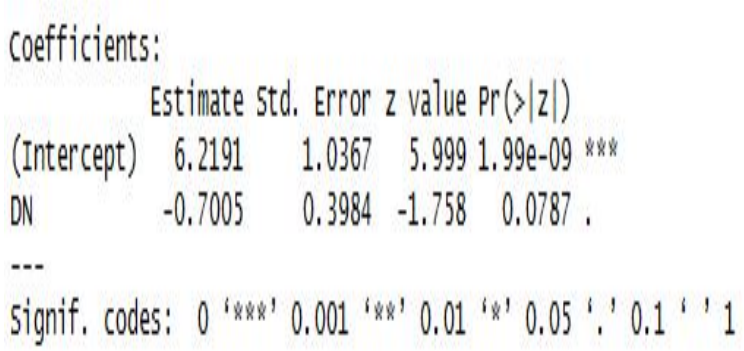


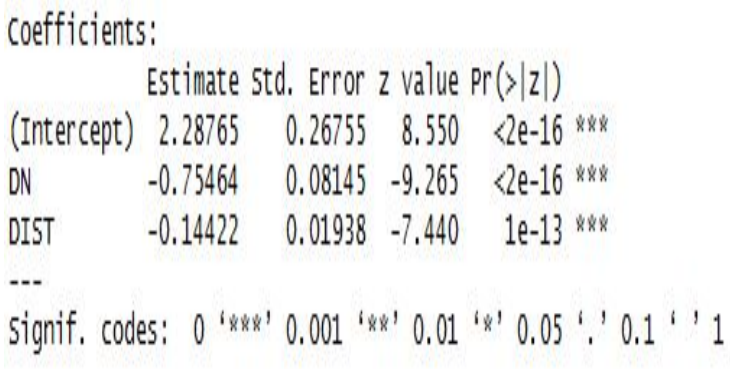
Figure 6: This is the output from R for the logistic regression with down as the predictor on the test set.

We now run the logistic regression with down and distance predicting play type on our test set. The output from R for our logistic regression can be seen in Figure 8. When we look over the results, we see that the down variable has a higher p-value but is still potentially considered significant, and that distance no longer has a p-value that allow the predictor to be considered significant. We've seen a similar situation to this in our linear regression setting, in which the training set was significant but the test set was not. Recall that in order to get around that we combined the sets and then ran the regression again and we were able to conclude the significance of our variables. We attempt the same strategy here by running the logistic regression on both the training and test set. The results of the regression can be seen in Figure 9. Looking over the results we see that the down variable has an improved p-value while the distance variable has p-value that does not improve. This suggests that for the test set, we can not conclude that down and distance predicted play type for the 2014 season.

With our findings the regression suggests that the play-calling in terms of predicting play type was not very predictable for 2014 season. If we consider the findings in our training a set a bit more, we can theorize that predictable tendencies from earlier seasons may have had an effect on the success of 2014 season. In other words the 2014 season may have not been predictable because defenses were able to apply strategies based of predictable tendencies in earlier seasons such that the offense changed play-calling style that lead to a lower amount of success.

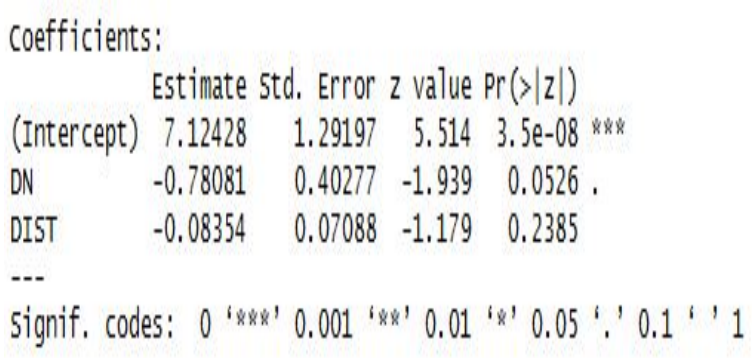
7.3 Personnel Predicting Play Type

Of all of the regressions ran so far, most are comprehensible in that the terms being used can be understood by the average watcher of football. This case is a bit different though, the term personnel is a term commonly used among football players and coaches to understand formations. In our case personnel is referring to the formations of Simpson's offense. Examples of personnel for formations include, 10, 22, 12, as well as many others. As you can see personnel involves a two digit numbering system. The first number is the number



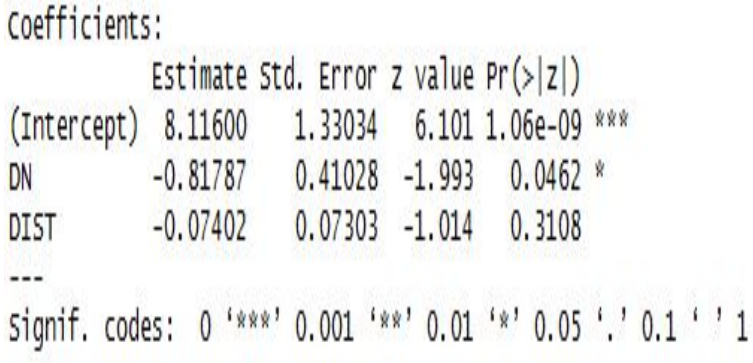
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.28765    0.26755   8.550   <2e-16 ***
DN          -0.75464    0.08145  -9.265   <2e-16 ***
DIST        -0.14422    0.01938  -7.440    1e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 7: This is the output from R for the logistic regression with down and distance as the predictors on the training set.



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.12428    1.29197   5.514  3.5e-08 ***
DN          -0.78081    0.40277  -1.939   0.0526 .
DIST        -0.08354    0.07088  -1.179   0.2385
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 8: This is the output from R for the logistic regression with down and distance as the predictors on the test set.



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.11600    1.33034   6.101 1.06e-09 ***
DN          -0.81787    0.41028  -1.993   0.0462 *
DIST        -0.07402    0.07303  -1.014   0.3108
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9: This is the output from R for the logistic regression with down and distance as the predictors on both the training and test sets.

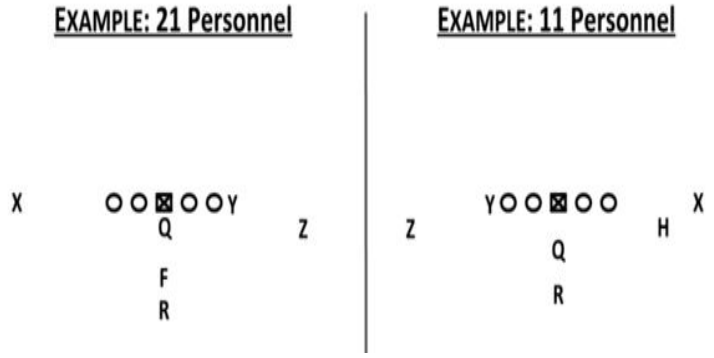
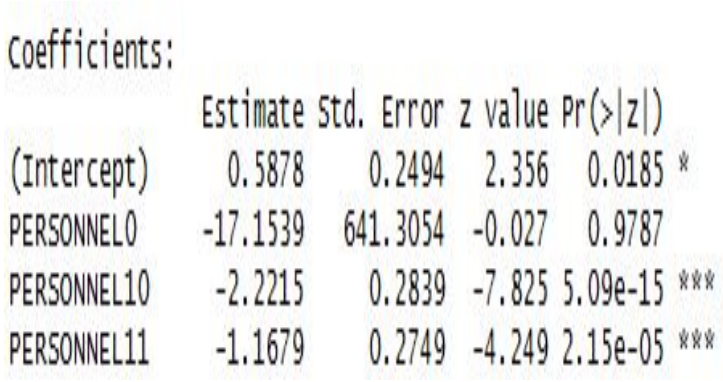


Figure 10: These are examples of formations for 21 personnel and 11 personnel.

of running backs involved in the formation. For the purpose of this project a running back is a player in the backfield within the tackle-box that is not the quarterback. Whereas a tight end is a player on the line of scrimmage, that is considered pass eligible. Examples of 21 and 11 personnel can be seen in Figure 10. Note that in each example diagram Y represents a tight end, R represents a running back, and F represents a fullback. There exist other players on each diagram, but for the purposes of determining personnel we are only concerned with tight ends and running backs.

Before we look at the regression analysis, we can look into more background why it can be beneficial to find the correlation between personnel and play type. Often times the formation of the offense can give the defense a hint on the play type that will be ran. For example if the offense is in a formation in which there are no true receivers, it could be expected that the play type will be run since the offense is better equipped to run the ball in that situation. This brings up the point of why are we looking at personnel rather than formation. The reason is that personnel categorizes the formations based on the number of tight ends and running backs in the formation. Therefore personnel encompasses multiple formations rather than looking at individual formations. There's a disadvantage in only analyzing individual formations in the regression setting because, some formations are solely predictable on their own without analysis. For example look at what is known as the victory formation. This formation is generally used when a team is winning toward the very end of the game, and there's no need to try and gain more yards. So out of the victory formation the quarterback simply takes a knee to run time off the clock. Thus in that case run can be predicted very easily. There exist other formations within offenses designed for specific plays, and since we want predictions on things we don't already know, it can be better to analyze multiple formations at once through the use of personnel.

Now that we have a good background on what personnel is, we can begin the logistic regression analysis. Before we look at the results from R, note that the full output will not be shown since there are 15 different personnel categories. Therefore only the useful results will be shown. The output for logistic regres-



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.5878     0.2494   2.356   0.0185 *
PERSONNEL0   -17.1539   641.3054  -0.027   0.9787
PERSONNEL10   -2.2215     0.2839  -7.825 5.09e-15 ***
PERSONNEL11   -1.1679     0.2749  -4.249 2.15e-05 ***
```

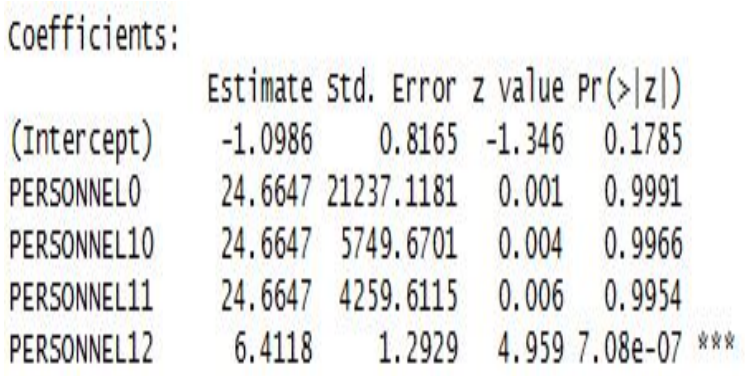
Figure 11: This is the output for the logistic regression with personnel predicting play type on the training set.

sion with personnel predicting play type on the training set can be seen in Figure 11. We see that both 10 and 11 personnel are considered significant based on their p-values in the output. Compared to our other logistic regressions these results are little more difficult to interpret. To determine the probability of pass or run for the personnel we substitute a 1 in for the x value associated with variable in the logistic function. To focus on one personnel grouping at a time we substitute 0 in for the other x values while we focus on the desired personnel grouping. In both 10 and 11 personnel situations, it predicted by the regression that both are more likely to be pass play with 10 personnel having about 0.83 probability of passing and 11 personnel having about 0.64 probability of passing. A possible explanation for the likelihood of passing is that there are not many running backs in either formation, and therefore the offense is better equipped to throw the ball.

We now run the logistic regression on our test set. The output from the regression can be seen in Figure 12. After some immediate analysis we see that both 10 and 11 personnel are no longer significant, but we see that 12 personnel is considered significant. There are some issues that arise similar to our logistic regression involving down and distance. The predicted probability of play type in this situation is 0.995 in favor of a run play. So similar to our most recent logistic regression, the model may not be a good fit for the test set.

8 Future Work

It's important to note in all of this analysis that there was only one classification or qualitative analysis done for this project. While linear regression can give useful outputs, we have seen through the analysis of data in the application that it may not always be the best model. Other classification methods to consider include linear discriminant analysis, quadratic discriminant analysis, and k-nearest neighbors. These methods may provide more insight for prediction, but each may have issues as well. A potential issue of each model is that while accuracy of prediction goes up, the interpretability of the model goes down. In



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.0986     0.8165  -1.346   0.1785
PERSONNEL0    24.6647 21237.1181   0.001   0.9991
PERSONNEL10   24.6647  5749.6701   0.004   0.9966
PERSONNEL11   24.6647  4259.6115   0.006   0.9954
PERSONNEL12    6.4118     1.2929   4.959 7.08e-07 ***
```

Figure 12: This is the output for the logistic regression with personnel predicting play type on the test set.

other words each model may be highly successful, but the results may be very difficult to interpret.

9 Conclusion

As we have seen, there are many items of item of information that suggest predictable play-calling may have occurred during the 2012-2014 seasons. For logistic regression specifically, it appears that the play type of any given play could have been be predicted with better probability than simply guessing during the 2012 and 2013 seasons. Given that the goal was to show that decline in success was due to something other than inexperienced players, it is somewhat difficult to pull a conclusion from the data analysis done the test set during this project. The models used in this project do not show a lot of solid evidence suggesting predictable play-calling during the 2014 season. Perhaps the previous seasons were predictable in way that the 2014 was not able to be successful because teams already had an idea of what was coming. It also very difficult to isolate one issue in a losing season, in other words athletic teams have many moving parts to them and it's difficult to determine what could be the sources contributing to a decline in success are. Maybe the previous teams were simply more talented and were able to overcome predictability, because the teams were so skilled. Then once new players were leading operations the predictability became more evident, and coaches were required to change their game plan which could lead to a decline in success, since changes were made mid-season. If the game plan worked in previous seasons, it would be reasonable to think it would work in the 2014 season, but a change in game plan would create a new learning curve. While the goal was to get away from the excuse of injuries, maybe injuries are what caused the change in game plan, and thus we saw less predictability. While it may have not been as successful as desired, perhaps this project has shown a different perspective on what could cause a decline in success.

10 Works Cited

- simpsonathletics.com
- The Simpson College football playbook
- Hudl.com
- Data for this project was made accessible by Ted Haag